Effects of dyadic work organization on realism in confidence judgments

Carl Martin Allwood
Department of Psychology, Lund University, Sweden

Pär Anders Granhag
and
Leif A. Strömwall
Department of Psychology, Göteborg University, Sweden

This study examines the effect of three regimens (organizations of activities) on the realism in dyad's confidence judgments. The effect of letting the pair members first work alone when selecting an answer and/or when confidence rating the correctness of this answer was compared with working directly in pairs. No improvement in the realism of the dyads' final confidence judgments was found as an effect of the work regimens involving preparatory work, compared with pairs having no such preparation. In fact, carrying out a confidence judgment acting as an individual after a dyadic decision about the selection of answer alternative, led to increased overconfidence in the pairs. Improved realism was found when pair members worked individually when selecting and confidence judging the answer, before collectively deciding on an answer but again giving individual confidence ratings, but this effect did not reach significance. It is suggested that these results might at least partly be explained by increased confidence as an effect of explaining one's belief. Finally, each pair's final judgment of the total number of questions answered correctly was not improved by individual preparatory work.

Key words: Dyads, confidence judgments, realism, individual reflection, metacognition.

Acknowledgements: Correspondence concerning this paper should be addressed to: C.M. Allwood, Dept. of Psychology, Lund University, Box 213, SE - 221 00 Lund, Sweden. E-mail: Cma@psychology.lu.se. This research was supported by a grant to the first author from the Swedish Council for Research in the Humanities and the Social Sciences. Thanks to Marcus Johansson for valuable comments on a previous version of this manuscript and to Anna-Carin Jonsson for help with collecting the data.

Discussing and counselling are common human activities. Investigating a large group of university students, Heath and Gonzalez (1995) showed that as many as 91% indicated that they sought the advice of others while making important decisions. On such occasions arguments based on assumptions about facts (implicitly or explicitly asserted) are put forth and attempts may be made to establish facts about the world. Moreover, the extent to which a suggested fact holds may be questioned, and the involved parties may attempt to judge how certain they are that the suggested fact actually holds. In other words, they may judge their confidence in the correctness of the suggested fact. In these contexts conclusions are commonly drawn both during ongoing social interactions and when individuals reflect on their own before, or after, the social interaction.

The purpose of the present study was to improve our understanding of factors affecting the metacognitive judgments made by pairs (dyads). More specifically, we investigated how the realism in confidence judgments is influenced by the fact that they were made by pairs working together and how the pair's realism is affected by independent individual reflection. By realistic confidence judgments we mean that there is, over many occasions, a balance between the confidence level used and the accuracy level of the items judged at that confidence level. As an example, perfect realism (calibration) implies that 60% of the items judged as 60 % sure should be correct, and that 100% of the items judged as 100 % sure should be correct. A common finding in calibration research has been that the participants show overconfidence in their answers to general knowledge questions (Allwood & Granhag, 1999; Lichtenstein, Fischhoff & Phillips, 1982; McClelland & Bolgar, 1994).

Previous research comparing the performance of groups and individuals on cognitive tasks has usually examined only the accuracy aspect (for reviews, see Gigone & Hastie, 1997; Hinsz, Tindale & Vollrath, 1997) or, more in general, the selection of one of a number of alternatives (Kerr, MacCoun & Kramer, 1996). These studies have commonly used within-subjects designs, and the most frequent result has been that the accuracy level of groups is higher than the mean of the individuals' accuracy level, but that the group does not outperform the accuracy of its best member (Gigone & Hastie, 1997).

Previous research has also found that certain aspects of the social interaction can affect the group's performance in a negative way. For example, the outcome of brainstorming is negatively affected by the fact that the other group members' speech activity interrupt or hinder the thought process of the individual members and thereby lower their productivity (Diehl & Stroebe, 1987, cited in Wilke & Kaplan, 2001). Along similar lines, Andersson and Rönnberg (1997), comparing the memory performance of singles and dyads, found evidence that inappropriate cueing between the pair members resulted in lower memory performance in the pairs.

Recently, researchers have begun to show some interest in the social aspects of metacognitive performance and how such aspects influence the validity of metacognitive judgments (e.g., Yzerbyt, Lories, & Dardenne, 1998). In the calibration context, certain features of social interactions have been found to result in poor realism in confidence judgments. Allwood and Granhag (1996a) found that when one member in a pair dominated the interaction by being identified by the dyad as the knowledgeable member in the area considered poor realism resulted. Likewise, poor realism resulted when all that happened in the interaction was that one member provided an argument in favor of the chosen alternative. As will be described below, research has shown that in the context of realism in confidence judgments such negative features of dyadic interaction can be compensated by combining the dyadic interaction with opportunity for individual reflection in the process.

Not much research has compared individuals with dyads (or larger groups) with respect to the confidence with which they hold their performance to be correct (e.g., Sniezek & Henry, 1989; Stephenson & Wagner, 1989). Using a within-subject design, Sniezek and Henry (1989) asked their participants, acting as individuals and in triads, to give 99% confidence intervals for the frequency of each of 15 causes of death in the US population. The triads had higher accuracy and gave more narrow confidence intervals than the individuals, which was interpreted as showing that the triads were more confident than the individuals.

Stephenson and Wagner (1989) first let their participants witness an event that they later, individually or in pairs, reported on by means of free-recall. Finally, the participants, again individually or in pairs, answered a number of questions about the event and confidence rated their answers. The pairs were more accurate and gave higher confidence ratings than the individuals. However, in none of these studies calibration measures were computed and it is thus hard to draw conclusions about the realism in the confidence judgments given by the pairs and the individuals.

To our knowledge, only Allwood and Björhag (1990), Allwood and Granhag (1996a) and Allwood, Granhag and Johansson (2003) have used calibration methods (described in the Method section, below) when investigating potential differences between pairs and individuals in the realism in confidence. Allwood and Björhag (1990) compared the realism of singles' and dyads' confidence judgments to general knowledge questions using a between-subjects design. No differences in accuracy were found, but the pairs showed significantly higher confidence compared with the individuals. Finally, no differences in the realism of the confidence judgments were found.

Allwood and Granhag (1996a) used a within-subject design where the participants first individually answered and confidence rated a set of general knowledge questions. Next, for each of a new set of questions answered, the participants individually gave an argument for the chosen answer and then confidence rated their certainty in the answer. Finally, the participants were paired into dyads and answered and confidence rated their answers to the second set of questions encountered. The main result was that overconfidence *decreased* for the pairs as compared with the two measures of the singles' results for the two sets of questions. In a control study, Allwood and Granhag (1996a) found that there was no effect of asking individuals to give confidence judgments twice to the same questions.

Allwood et al. (2003), showed a short film and the participants in one condition (the Individual-Pair condition) answered and confidence rated questions on the film, first individually and then in pairs. In another condition (the Pair condition) the participants only answered the questions in pairs. The pairs in the Individual-Pair condition had better realism in their confidence judgments as measured by calibration than the pairs in the Pair condition. Furthermore, a comparison within the Individual-Pair condition replicated the result reported by Allwood and Granhag (1996a), i.e., the participants showed less overconfidence when they acted as a pair, compared with when they acted as individuals. Finally, in a control study Allwood et al. (2003), again, in similarity to Allwood and Granhag (1996a), found no effect on realism of asking individuals to give confidence judgments twice to the same questions.

In brief, the reviewed research shows that the regimen (organization of activities) of first letting the participants individually answer the questions and confidence rate their own answers, and then to collaborate on the same tasks in dyads, can be a successful means of improving dyadic realism. However, it is not clear what parts of the individuals' activities improved the pairs' realism. In order to better our understanding of the processes, the present study investigated the effect of letting the participants reflect individually on their confidence judgments after they had answered the questions collectively together. More specifically, we investigated the possibility of improving realism in pairs' confidence judgments by introducing individual metacognitive reflection, i.e., individual reflection on the appropriate level of the confidence judgments after the knowledge question had been collectively answered, but before the pair agreed on the confidence judgments.

Generally speaking, good realism in confidence judgments is dependent on the relation between accuracy and confidence. Simply put, if a person is overconfident, an increase in accuracy and/or a decrease in confidence will lead to improved realism. Different theories have been suggested for explaining the degree of realism in confidence judgments given by individuals of their answers to general knowledge questions. The most important of these theories were reviewed by McClelland and Bolgar (1994). A common feature of many of these theories is the idea that it is the thought content activated when the answer alternative is selected that has a decisive influence on the confidence level reported (e.g. Allwood & Granhag, 1996b; Gigerenzer, Hoffrage & Kleinbölting, 1991; Koehler, 1994). This either suggests that no (or little) new content is generated when the confidence judgments are made, or that this content is of less importance for the level of the confidence judgments than the thoughts generated when the knowledge question is answered.

Observations reported by Allwood and Björhag (1990) and by Allwood and Granhag (1996a) support this conclusion for dyads. In both these studies, very little negotiation or other interaction took place within the dyad when the level of the confidence judgment was decided. Our impression from these two studies was that, when the confidence judgment was to be given, it was already, to a large extent, 'in the air'. Similarly, Stephenson and Wagner (1989) reported that far more time was spent discussing facts relating to answering the actual question than discussing the level of the confidence judgment.

If it is the case that the thoughts generated when the knowledge question is answered are more important than the unique thoughts generated when making the confidence judgment, one contributing reason for this may be that the social context is the same both when the knowledge question is answered and when the confidence judgment is given. Thus, the stability in the social environment might limit any new thinking when the confidence judgment is made. However, if the individuals in the pair are allowed to reflect individually on the confidence judgment before collectively agreeing on the confidence level, this may lead to new information being introduced to a greater extent when the pair's confidence judgment is made. This, in turn, might lead to greater difference of opinions and, as an effect of this, to more realistic confidence judgments. More specifically, taking into consideration that previous research shows that overconfidence can be expected in the pairs' confidence judgments, the greater amount of discord that might result in the pair after individual reflection on the confidence judgments might be expected to contribute to a decrease in overconfidence.

Since individual reflection on the confidence judgments before they are given by the pair may not be sufficient to improve the pairs' realism, we, in one condition attempted to further improve the pairs' realism by first, before the knowledge questions were answered by the pair, letting the participants also reflect on the knowledge questions as individuals. By letting the participants first consider the questions acting as individuals, we aimed at enabling the pair members to contribute to better knowledge utilization in the dyadic discussion. This was expected to lead to an increase in the pairs' accuracy that would *per se* contribute to improved realism in the following confidence judgments. However, the differences in the pair members' experiences when they individually considered the knowledge questions was also expected to contribute to differences in opinions and more realistic confidence ratings by the pairs.

Effects of repeated consideration of knowledge questions

In two of the three conditions in the present experiment the participants deal with the same knowledge questions more than once. Previous research has suggested that repeated answering and confidence rating can *per se* influence the level of the confidence judgments made. In fact, two different repetition effects have been reported in the literature. In one type of repetition effect, what matters is the number of times the participant states the same answer to the same question. Hertwig, Gigerenzer and Hoffrage (1997) reviewed research on this first type of effect, stemming from the memory and decision-making literature, and called it the *reiteration effect* (see also, Koehler, 1991).

The second type of repetition effect was reported by Granhag (1997), in a study where the participants confidence rated *the same* answers on two occasions. In contrast to the results relating to the reiteration effect discussed by Hertwig et al., the results showed that the confidence level was *lower* when the participants repeated their confidence judgment of their answer to the same question. This result was replicated by Granhag, Strömwall and Allwood (2000). In both these studies the duration between the two occasions when the confidence judgments were made was one week or more.

Frequency judgments

Finally, at the end of each of the three conditions studied, we asked the participants to give an estimate of how many of all the questions answered they believed that they (when acting as a pair) had answered correctly. In other words, they were asked to perform what will henceforth be called a *frequency judgment*. To our knowledge, only Allwood et al. (2003) has investigated pairs carrying out frequency judgments. The results showed that the pairs either underestimated their actual performance or were realistic in their frequency judgments.

Previous research focusing on individuals and general knowledge questions or questions probing episodic memory has shown that frequency judgments tend to show good realism or to underestimate the number of correct items. When the participants' item specific confidence judgments show good realism, their frequency judgments tend to underestimate the number of correct items (Gigerenzer et al., 1991) and when the participants' item specific confidence judgments show overconfidence, their frequency judgments tend to show rather good realism (Allwood & Granhag, 1996b; Granhag et al., 2000; Treadwell & Nelson, 1996, exp 1, and previous research cited in that paper). One study reported over-optimistic assessment in the frequency judgments (Treadwell & Nelson, 1996, exp 2).

Allwood and Granhag (1996b) and Granhag et al. (2000) attempted to explain these results by initially noting that frequency judgments, compared with item specific confidence judgments, are made at a greater temporal distance from the target of the judgment. For this reason the participant was assumed to heed a greater variety of information about various aspects of the target or about themselves as epistemological subjects. This, in turn, may lead to a decrease in the estimated number of correct items. This explanation is similar to the explanations given by other researchers reviewed by Treadwell and Nelson (1996) under the label "dual source" explanations. These explanations have in common that different types of information are used for the two types of judgments.

Hypoteses

We had four hypotheses. The first was that we expected more realistic confidence judgments in the two conditions where the pair members first reflected individually on their confidence judgments compared with the condition where the participants acted directly as a dyad. The reason for this hypothesis was that we expected that individual reflection would lead to an increase in the number of unique associations later available to the pair when they discussed the confidence level for the same questions. We reasoned that consideration of a broader range of contents would lead to more realistic confidence judgments.

The second hypothesis related to the two conditions where the participants did not act directly as a pair. The hypothesis was that the pair members' prior individual confidence judgments would lead to an improvement in realism of the confidence judgment they made together. Our reason for this hypothesis was the same as for the first hypothesis.

The first part of our third hypothesis stated that the pairs' accuracy would increase in the condition where they first answered the knowledge questions individually and the second part of the hypothesis was that the increase in accuracy would be associated with an improvement in our measures of realism. In similarity to our first hypothesis, our reason was that individual consideration of the questions would lead to better knowledge utilization in the pair's discussion.

Finally, our fourth hypothesis stated that the participants, in line with prior research would show overconfidence in their item specific confidence judgments and fairly realistic frequency judgments. However, since previous research has suggested that item-specific confidence judgments and frequency judgments may be influenced by different processes and that the processes affecting frequency judgments may be more general in kind, we did not hypothesize any differences between the three conditions in this context.

Method

Participants

A total of 120 undergraduate students (79 women, 41 men, mean age 25 years) from Göteborg University volunteered to participate in the experiment and were paid 100 SEK for their efforts.

Design

The participants were initially randomly teamed up in pairs. Each pair was randomly allocated to one of the three experimental conditions (Simple pair, Repeat 1, and Repeat 2). In the Repeat 1 and Repeat 2 conditions, the participants provided repeated assessments of confidence. This made it possible to undertake within-subjects comparisons within these conditions, as well as between-subjects analyses comparing the three conditions.

Materials

All participants answered, and confidence rated, the same 80 general knowledge questions on paper. For each question, a forced choice, two-answer alternative, response mode was employed. The questions have been used in previous research (Allwood & Granhag 1996b), and proved not to be too easy (90% or more correct answers) or too difficult (less than 50% correct answers). The questions covered topics such as history, geography, literature, society, and linguistics. The confidence judgments were always given on a scale from 50% to 100%, where it was explained to the participants that 50% meant that they were absolutely unsure as to the correctness of their chosen answer alternative ("guessing"), and 100% indicated absolute certainty in the correctness of the alternative chosen. It was instructed that these endpoints as well as any numeral between these endpoints could be used in stating the confidence judgments and that one of the two answer alternatives was always correct.

Procedure

Table 1 provides an overview of the different phases within each of the three experimental conditions.

Simple pair condition. This condition had only one phase. The pairs were instructed to collectively discuss and try to select the correct answer and then to give a confidence judgment for each question. For each question, the pair first selected an answer and then immediately confidence rated their choice. The pair members were instructed that the answer alternatives chosen, and the confidence levels expressed were supposed to be the product of a thorough discussion, not just allowing one pair member to take full control.

Repeat 1 condition. In Phase 1 in this condition, the pairs were first instructed to answer each question together, that is, they had to come to an agreement regarding which answer was correct. Both pair members marked this agreed-upon alternative on their own separate answer sheets. Next, directly after each question, the confidence in the selected answer was provided on an individual basis. The participants were instructed to write down their individual confidence judgment without discussing or showing this numeral to the other pair member. This was done for all 80 questions.

Table 1

Overview of the Experimental Procedure

Condition	Phase	Description of Each Phase
Simple pair	1	The pair members made collective answers and collective confidence judgments.
Repeat 1	1	The pair members made collective answers, but individual confidence judgments.
	2	The pair members made collective confidence judgments.
Repeat 2	1	The pair members made individual answers and individual confidence judgments.
	2	The pair members made collective answers, but individual confidence judgments.
	3	The pair members made collective confidence judgments.

In Phase 2, the pair members were asked to come to an agreement regarding the confidence judgment for each of the answers selected in Phase 1. They were explicitly told that they should make an effort in discussing this collective confidence judgment, not just mechanically take the average of their separate numerals. They were not allowed to change the chosen answer alternatives.

Repeat 2 condition. In Phase 1, the participants first individually answered and confidence judged each of the 80 questions. The next two phases were identical to Phase 1 and Phase 2 in the Repeat 1 condition, described above.

Finally, the pairs, in all conditions, made a frequency judgment; that is, they were asked to make an overall assessment, without looking back at the answer sheets, as to the number of correct answers they had provided in their collectively agreed answers. The participants were then debriefed and paid for their efforts.

The experiment lasted, on average, about 50 minutes (Simple pair condition), 70 mins (Repeat 1 condition), and 90 mins (Repeat 2 condition). All interactions between the pair members were audio-recorded, with their consent. The analysis of the recordings is presented in a separate paper (Johansson, Allwood & Granhag, 2002).

Dependent measures

We used two measures, derived from the Brier score formula, to analyze the realism in participants' confidence judgments: calibration and over/underconfidence. Both measures and their formulae are further described in Lichtenstein, Fischhoff and Phillips (1982). *Calibration* reflects the overall relation between the level of the confidence judgments and the accuracy. A calibration score of 0 indicates perfect calibration, and the higher the calibration value the poorer the realism. The formula for computing calibration is:

$$(1) \qquad \text{Calibration} = 1/n \sum_{t=1}^{T} n_t \, (r_{tm} - c_t)^2$$

In (1), n is the total number of questions answered, T is the number of confidence classes used, c_t is the proportion correct for all items in the confidence class r_t, n_t is the number of times the confidence class r_t was used and r_{tm} is the mean of the confidence ratings in confidence class r_t. Thus, calibration is computed by first dividing participants' confidence ratings into a number of confidence classes. Next, for each confidence class, the difference is taken between the mean confidence for the items and the proportion of correct items. Finally, the squared differences multiplied by the number of responses in the confidence class are summed over confidence classes and divided by the total number of items.

The *over/underconfidence* measure (henceforth called overconfidence) indicates whether a rater is overconfident (positive value) or underconfident (negative value). Overconfidence is computed in the same way as calibration, except that the differences are not squared. Higher absolute over/underconfidence value indicates higher over- or underconfidence, or less realistic confidence judgments. A value of zero indicates neither over- nor underconfidence.

Results

In the results section we first present a between-subjects analysis, comparing the three conditions on the dependent measures. Next follows a within-subjects examination of the two Repeat conditions. Finally, we report the results of the frequency judgments analyses.

Between-condition analyses

At the end of each condition, each pair had together answered each question and confidence judged each answer. Calibration curves for the results of the last phase of each of the three conditions are shown in Figure 1. These calibration curves show the proportion of correct answers and the mean confidence for each confidence class (50–59, 60–69, 70–79, 80–89, 90–99, 100). As can be seen in Figure 1, each condition showed some underconfidence at the lowest level (50–59), indicated by their mean accuracy being above the reference line. For all other confidence classes, overconfidence was found. Finally, Figure 1 also shows that the three lines representing the conditions follow each other closely.

The means and standard deviations for the dependent variables in the last phase of each condition are presented in Table 2[1]. In all conditions, the pairs showed overconfidence, varying from .07 to .11. Four one-way ANOVAs were computed in order to compare differences between the three conditions, one for each of the dependent measures. No significant effects were found. The different work regimens did not, in their final phases, lead to any significant differences in the realism of the confidence judgments.

Within-conditions analyses

To trace the effects of the phases (work regimens) the pair members went through, we analyzed the Repeat 1 and Repeat 2 conditions separately. The descriptive statistics for these

[1] We also computed the results for *resolution*. This measure reflects the ability of the participant to distinguish as clearly as possible between two sets of answers, one set that is correct and one set that is incorrect, by means of the confidence judgments. The formula for the resolution measure is given in Lichtenstein, Fischhoff and Phillips (1982). Since no theoretically interesting differences were found between or within the conditions these results are not included.

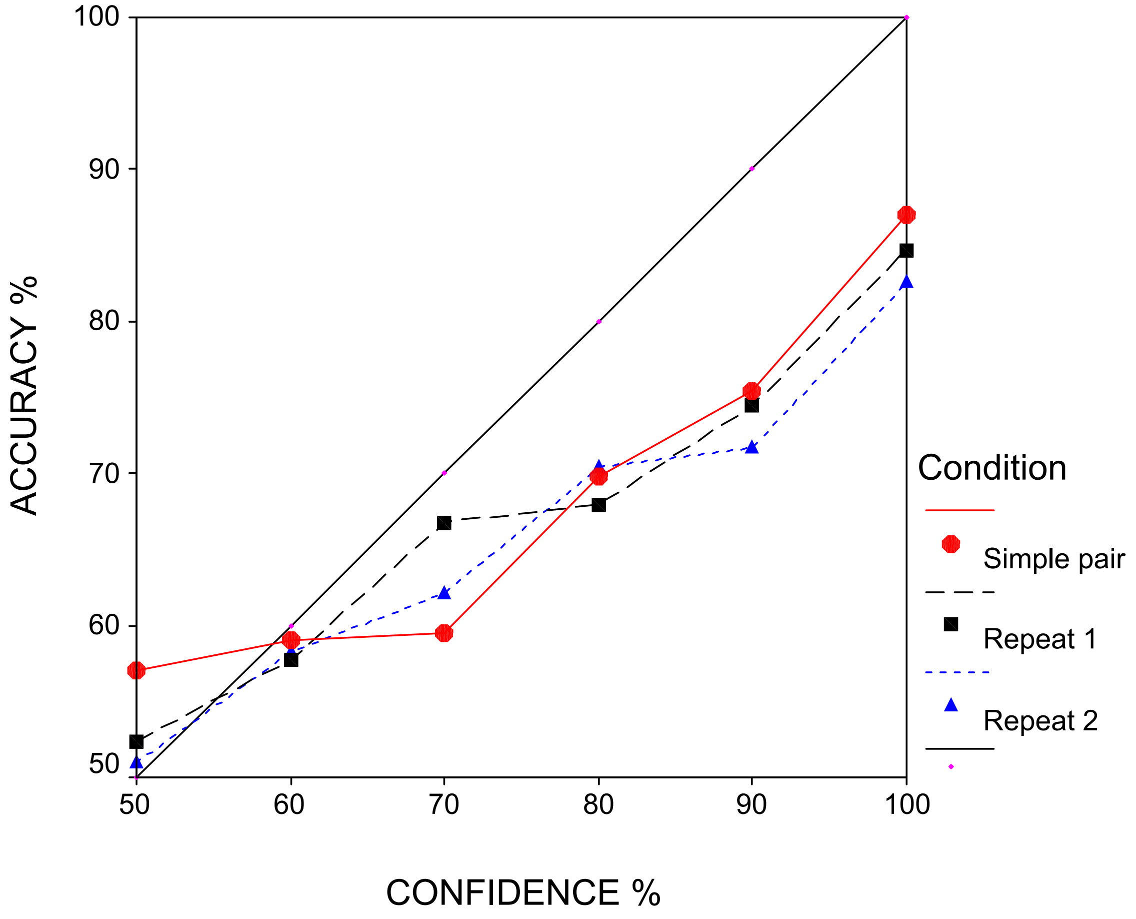

Figure 1. Calibration curves for the results of the last phase of each of the three conditions.

within-subjects comparisons are given in Table 3 for the Repeat 1 condition, and in Table 4 for the Repeat 2 condition. For the first (Repeat 1 condition), and the first and second (Repeat 2 condition), phase in these two conditions the results are presented in three ways. First, the results are shown for the mean of the members in each pair with the smallest value (min), second, for the mean of the average of the two pair members (average), and third, for the mean of the members in each pair with the largest value (max). The min and max values were included in order to examine the extent to which the pairs' performance exceeded the members with the lowest and the highest values. These values also gave some overall indication which of the pair members, the one with the highest or the lowest initial individual value, influenced the subsequent collective confidence judgment the most. In order to examine differences in the pairs' performance over time, pairwise *t*-tests were run.

Repeat 1 condition. The descriptive statistics are presented in Table 3. Note that the comparisons within this condition concern possible differences in the participants' realism in confidence when they performed individual and collective confidence judgments. Also note that they could not alter their answers to the general knowledge items between the two phases. Hence, accuracy could not change between the occasions, and any difference in mean confidence is also a difference in the over/underconfidence measure.

Table 2

Means (SD's in parentheses) for the Dependent Measures by Condition and Results of Between-subjects Comparisons

	Condition				
Measure	Simple Pair	Repeat 1 Phase 2	Repeat 2 Phase 3	*F*	*p*
Calibration	.0306 (.0228)	.0292 (.0206)	.0324 (.0260)	<1	.907
Over/Underconfidence	.0728 (.0948)	.0935 (.0921)	.1079 (.0818)	<1	.467
Accuracy	.6797 (.0878)	.6769 (.0722)	.6706 (.0795)	<1	.935
Confidence	.7526 (.0829)	.7703 (.0735)	.7785 (.0597)	<1	.518

Note. The *F*- and *p*-values refer to one-way ANOVAs with df = 2, 57.

For *calibration* both the min and the max values (but not the average value) in Phase 1 differed significantly from the pair's collective calibration in Phase 2. The collective calibration in Phase 2 was worse compared with the member with the best (min) calibration (pairwise *t*(19) = -2.67, *p* < .05.) and better compared with the member with the worst (max) calibration (*t*(19) = 3.48, *p* < .01).

For *overconfidence* (and *confidence*) the min and the average values in Phase 1 were significantly lower (i.e., better) compared with the pair's collective overconfidence in Phase 2, pairwise *t*(19) = -4.98, *p* < .001, and pairwise *t*(19) = -3.90, *p* < .01, respectively. The max value in Phase 1 was significantly higher (i.e., worse) than the collective overconfidence in Phase 2, pairwise *t*(19) = 2.71, *p* < .05. Since there was a smaller difference between the max value (Phase 1) and the pairs' value (Phase 2) than between the min value (Phase 1) and pairs' value (Phase 2), the results suggest that it was the pair members expressing the highest confidence in the individual Phase 1 who influenced the subsequent, collective confidence judgment more.

Table 3

Means (and Standard Deviations) for the Two Phases of the Repeat 1 Condition (n = 20)

	Phase 1			Phase 2
Measure	Min	Average	Max	
Calibration	.0216* (.0172)	.0311 (.0183)	.0406** (.0223)	.0292 (.0206)
Over/Underconfidence	.0504*** (.0928)	.0780** (.0902)	.1055* (.0944)	.0935 (.0921)
Accuracy	–	.6769 (.0722)	–	.6769 (.0722)
Confidence	.7273*** (.0776)	.7548** (.0736)	.7824* (.0780)	.7703 (.0735)

Note. Probability asterisks indicate a significant difference compared with Phase 2.

* *p* < .05, ** *p* < .01, *** *p* < .001.

Repeat 2 condition: comparisons between Phase 1 and 2. For the Repeat 2 condition, the first pairwise comparisons concern differences between the min, average and max values of each of the dependent measures in Phase 1 on the one hand, and the average of the same dependent measures in Phase 2 on the other. Descriptive statistics for the dependent variables in each of the three phases of the Repeat 2 condition are given in Table 4.

For *calibration* both the min and the max values (but not the average value) in Phase 1 differed significantly from the average calibration in Phase 2. The average calibration in Phase 2 was worse compared with the member with the best (min) calibration in Phase 1, $t(19) = -2.40$, $p < .05$ and better compared with the member with the worst (max) calibration in Phase 1, $t(19) = 2.44$, $p < .05$. For *overconfidence* the max value in Phase 1 was significantly higher than the average overconfidence in Phase 2, $t(19) = 3.83$, $p < .01$.

For *accuracy* all three values (min, average, max) in Phase 1 were significantly lower than the average accuracy in Phase 2 (pairwise $t(19) = -8.87$, $p < .001$; pairwise $t(19) = -6,79$, $p < .001$; and pairwise $t(19) = -2.23$, $p < .05$, respectively). In other words, the pairs outperformed even the individual with the highest accuracy. Finally, for *confidence* both the min and the average confidence were lower in Phase 1 compared with the average confidence in Phase 2, pairwise $t(19) = -7.99$, $p < .001$, and pairwise $t(19) = -4.43$, $p < .001$, respectively.

Taken together, the statistical tests carried out show that while the accuracy did increase substantially over time, so too did the confidence levels. As a consequence, neither the calibration nor the over/underconfidence measures showed any significant improvement when moving from individual (Phase 1) to collective (Phase 2) answers.

Repeat 2 condition: comparisons between Phase 2 and 3. In this section, the dependent measures computed for the individually performed confidence judgments of the collective answers in Phase 2 are compared with same dependent measures for the collectively performed confidence judgments of the same answers in Phase 3 (see Table 4). These are the same comparisons as performed within the Repeat 1 condition above. Again, the accuracy could not change between the two phases, and any change in mean confidence is also a change in overconfidence.

For *calibration* both the min and the max values (but not the average value) differed significantly from the pair's collective calibration in Phase 3. The collective calibration in Phase 3 was worse compared with the member with the best (min) calibration in Phase 2, pairwise *t*(19) = -2.80, $p < .05$, and better compared with the member with the worst (max) calibration in Phase 2, pairwise *t*(19) = 5.06, $p < .001$.

Table 4

Within-subjects Comparisons for the Three Phases of the Repeat 2 Condition (n = 20)

	Phase 1			Phase 2			Phase 3
Measure	Min	Average	Max	Min	Average	Max	
Calibration	.0237*	.0387	.0538*	.0243$^{\#}$	.0342	.0441$^{\#\#\#}$	.0324
	(.0164)	(.0249)	(.0408)	(.0207)	(.0219)	(.0259)	(.0260)
Over/	.0482	.1006	.1530**	.0551$^{\#\#\#}$	.0839$^{\#\#\#}$	.1126	.1079
Underconfidence	(.0704)	(.0713)	(.0917)	(.0903)	(.0846)	(.0838)	(.0818)
Accuracy	.5569***	.6022***	.6475*	–	.6706	–	.6706
	(.0686)	(.0613)	(.0647)		(.0795)		(.0795)
Confidence	.6612***	.7028***	.7443	.7257$^{\#\#\#}$	.7545$^{\#\#\#}$	.7832	.7785
	(.0527)	(.0532)	(.0597)	(.0689)	(.0648)	(.0673)	(.0597)

Notes.

* indicates a significant difference between Phase 1 (min, average, or max) and Phase 2 (average).

$^{\#}$ indicates a significant difference between Phase 2 (min, average, or max) and Phase 3.

* $p < .05$, ** $p < .01$, *** $p < .001$; $^{\#}$ $p < .05$, $^{\#\#\#}$ $p < .001$.

For *overconfidence* (and *confidence*) both the min and the average overconfidence were lower (i.e., better) in Phase 2, compared with the collective overconfidence in Phase 3, pairwise *t*(19) = -7.97, $p < .001$, and pairwise *t*(19) = -7.41, $p < .001$, respectively.

Frequency judgments

At the end of the experiment all pairs, in all conditions, estimated their total number of correct answers. In order to facilitate comparisons, the frequency judgments were transformed into percentages of correct answers. Table 5 presents the results for the estimated and actual accuracy percentages. The mean frequency judgment percentages were compared between the three conditions, using a one-way ANOVA. No significant difference was found, $F(2, 57) = 1.65$, $p = .20$.

Table 5

Estimated and Actual Percentages (SD's) of Accuracy by Condition

Condition	Estimated %	Actual %	Difference	*t*	*p*
Simple Pair	64.31 (18.32)	67.97 (8.78)	-3.65	-1.04	.31
Repeat 1	72.81 (10.39)	67.69 (7.22)	5.12	1.91	.07
Repeat 2	70.31 (15.79)	67.06 (7.95)	3.25	1.07	.30

Note. The *p*-values refer to *t*-tests within each condition comparing the estimated and actual percentage of correct answers

We further compared, with pairwise *t*-tests, the estimated with the actual percentages of correct answers to examine if the work regimens led to differences in the realism of the frequency judgments. Table 5 contains the results of these analyses. Participants in the Simple pair condition underestimated their performance and participants in both the Repeat 1 and the Repeat 2 conditions overestimated their performance. None of the differences was, however, significant.

Discussion

In the present study we compared the effects on realism in dyadic confidence judgments of three different ways of organizing the work. In the Simple pair condition, constituting the most elementary organization, the pair both answered the general knowledge question and confidence judged the selected answer collectively. This condition was compared with two more complex work regimens, each constituting a separate condition (Repeat 1 and Repeat 2). In each of these conditions the pair members first individually answered and/or gave individual confidence judgments before delivering the collective performance, in different phases.

The main result of the study was that the two more elaborated work regimens did not lead to more realistic confidence judgments, compared with the Simple pair condition. This means that our first hypothesis was disconfirmed. It could be argued that the lack of effect was due to that the members in the dyads refrained from interacting when deciding upon their collective confidence judgment, and instead reached a rating by simply averaging their individual confidence judgments. In order to investigate this explanation we performed a content analysis sorting the activities within the dyads into four different interaction categories: Interaction, Dominance of one member, Immediate agreement and Immediate compromise (for a full presentation of this analysis, see Johansson, Allwood & Granhag, 2002). It was found that the dyads in the two Repeat conditions interacted significantly more (i.e., obtained a higher frequency for the Interaction category), when deciding on the collective confidence judgments, than did the dyads in the Simple Pair condition. Hence, introducing individual preparatory work did in fact lead to increased interaction. However, despite the more extensive cognitive work carried out by the participants in the Repeat 1 and Repeat 2 conditions, the end result, in terms of degree of realism in the confidence judgments, did not differ between the conditions. Thus, the extra individual reflection introduced concerning the appropriate level of the confidence judgments, before the pairs' collective confidence ratings, did not improve the realism of the pairs' confidence judgments. In fact, if anything, it appears to have increased the pairs' overconfidence.

We next discuss the results for the Repeat 1 condition and the results for the comparison between Phase 2 and 3 in the Repeat 2 condition. It is noteworthy that although accuracy could not improve in these comparisons (because of the design), the overconfidence increased significantly both between Phase 1 and 2 in the Repeat 1 condition and between Phases 2 and 3 in the Repeat 2 condition. Thus, our second hypothesis, stating that the pair members' prior individual confidence judgments would lead to an improvement in realism of the confidence judgment they made together was disconfirmed. It is possible that the individual with the highest confidence level (not necessarily the same person for all items) influenced the subsequent confidence judgments more. Reasonably, a person who is more convinced about something, for example that the chosen answer alternative is correct, tends to be more active and argumentative compared with a person who is less convinced about his or her opinion. This suggestion should be further investigated in future research.

The results for the Repeat 1 condition and for Phase 2 and 3 in the Repeat 2 condition can be compared to the results for the regimen studied by Allwood and Granhag (1996a). As described in the introduction, these authors first asked their participants to answer and confidence judge their answers to general knowledge questions individually. Next the individuals were teamed up into pairs and asked to answer and confidence judge both the same and different questions. The results showed less overconfidence in the confidence judgments by the pairs compared with the average overconfidence by the pair members when they made individual confidence judgments. Similar results were found by Allwood et al. (2003) in the context of episodic memory.

The difference in results between the present study and the studies by Allwood and Granhag (1996a) and Allwood et al. (2003) may be explained by the fact that the regimen investigated in the latter two studies initially provided each member of a pair with specific individual experiences *both* when answering the knowledge questions and when confidence rating the answers. In contrast, in the Repeat 1 condition and in Phase 2 and 3 in the Repeat 2 condition, in the present study, the pairs answered the questions together and only had specific individual experiences when giving the confidence judgments. In the introduction it was noted that many of the important theories on realism in confidence judgments assume that the realism in confidence judgments is mostly affected by the events occurring when the answer alternative is selected, and less by the events occurring when the confidence judgment is made. If this is the case, it seems reasonable that the regimen implemented in the Repeat 1 condition in the present study was less effective than the regimen implemented by Allwood and Granhag (1996a) and by Allwood et al (2003) due to the fact that it did not provide the two pair members with different experiences when the knowledge questions were answered.

We next discuss the results for Phase 1 and 2 in the Repeat 2 condition. Initially, it is noteworthy that the accuracy increased between the two phases even compared to the member with the highest accuracy in Phase 1. Thus, the first part of our hypothesis 3 received support. This result deviates from the more common research result showing that the accuracy of the group increases, compared with the average of the group members' individual accuracy, but not above the accuracy of the best individual (Gigone & Hastie, 1997). The fact that accuracy increased can be seen as evidence against the suggestion by Heath and Gonzalez (1995), that pair discussions do not result in information collection (or, as suggested by Allwood and Granhag 1996a, better "knowledge utilization") but only in construction of rationales.

However, the results also showed that the increase in confidence between Phase 1 and 2 in the Repeat 2 condition neutralized the possible beneficial effects of the increase in accuracy. Thereby the results contradicted the second part of our third hypothesis, although the results went in the expected direction since overconfidence decreased between phase 1 and 2 but not significantly so. It is not clear why the individual confidence judgments in Phase 2 did not lead to improved realism (at least not significantly) compared with Phase 1. However, a comparison with the studies by Allwood and Granhag (1996a) and by Allwood et al. (2003), where the pairs did in fact show a lower degree of overconfidence than the individuals, gives some clues as to the answer to this question.

An important difference between Phase 1 and 2 in the Repeat 2 condition in the present study and the studies by Allwood and Granhag (1996a) and by Allwood et al. (2003), is that the pair members in the latter studies made the confidence judgment together whereas this was not the case in phase 2 of the Repeat 2 condition in the present study. Given this difference, an explanation for the difference in results between the studies may be that any dampening effect of the pair's discussion, after the specific individual experiences had taken place in Phase 1 in the Repeat 2 condition, may have been lost, or "not taken home", since the two members were completely free to make whatever confidence judgment they felt like in Phase 2. For example, the member with the lowest confidence level could not, in Phase 2, provide a moderating influence on the pair member with the highest confidence level.

In addition, our finding that the confidence level of both the member with the lowest and the member with the highest confidence in Phase 1 went up in Phase 2 (not significantly for the member with the highest confidence) also suggests that the pair member who had been most confident in Phase 1 (not necessarily the same person for all items) appeared to have influenced the collective confidence judgment the most. This is in line with the notion suggested by Koehler (1991) according to which people's confidence increases when they explain their belief and the somewhat similar notion by Hertwig et al. (1997) that a person's confidence increases when they repeat their beliefs.

The reiteration effect presented by Hertwig et al. is one of the two types of repetition effects discussed in the introduction. The second was the repetition effect reported by Granhag (1997). In the first type of effect, confidence is expected to increase and in the second case it is expected to decrease. Our design is not an exact replication of the situations described by Hertwig et al. since the participants in our Repeat 2 condition were asked to take a stand once more with respect to whether or not their initial answer was correct. They were not simply asked to assert their answer one more time. However, our design is still further away from the situation studied by Granhag (1997) since the participants in Phase 2 of the Repeat 2 condition both (collectively) answered the questions again and confidence rated these answers (individually). Furthermore, the repetition of the confidence judgment took place very soon after the first judgment was given, not after a week as in the Granhag study. Thus, overall, we speculate that it was the reiteration effect of Hertwig et al. that was the most important for our results in the Repeat 2 condition, not the repetition effect reported by Granhag (1997).

Finally, we discuss the results for the frequency ratings. Just as for the item specific confidence judgments, the frequency judgments did not differ between the three conditions. For each of the three conditions the result followed the pattern shown in previous research, i.e., when the item specific confidence judgments show overconfidence, the frequency judgments show fairly good realism (Allwood & Granhag, 1996b; Treadwell & Nelson, 1996, exp 1). Thus, our results gave support to our fourth hypothesis. Moreover, the result of the current study, i.e., fairly realistic frequency judgments, support the finding of Allwood et al. (2003) - the only previous study investigating how dyads perform frequency judgments.

The fact that two different scales were used in the item specific judgments and the frequency judgments may not have been of decisive importance since Treadwell and Nelson (1996) found that the difference with respect to the two scales used only explained about 29% of the difference between the two types of judgments.

Much human cognitive activity occurs in social settings where social interaction and individual reflection are mixed. In the present study we attempted to improve pairs' metacognitive performance by introducing individual reflection on the appropriate level of the confidence judgments before they were given by the pairs. Our results show that it makes a difference to metacognitive performance if collaboration takes place or not. Different ways of organizing work gave different effects on the realism in confidence judgments. However, our results also show that for individual reflection to improve pairs' metacognitive realism, the individual reflection on the metacognitive issue only is not enough, it has to include individual reflection on the primary issue of which answer alternative is the correct one to choose.

References

Allwood, C.M., & Björhag, C.-G. (1990). Are two judges better than one? On the realism in confidence judgements by pairs and individuals. In J.-P. Caverni, J.-M Fabre, & M. Gonzalez (Eds.), *Cognitive Biases* (pp. 443-463). Amsterdam: Elsevier Science Publishers B.V. (North Holland Publishing Company).

Allwood, C.M., & Granhag, P.A. (1996a). Realism in confidence judgments as a function of working in dyads or alone. *Organizational Behavior and Human Decision Processes, 66*, 277-289.

Allwood, C.M., & Granhag, P.A. (1996b). Considering the knowledge you have: Effects on realism in confidence judgements. *The European Journal of Cognitive Psychology, 8*, 235-256.

Allwood, C.M., & Granhag, P.A. (1996c). The effects of arguments on realism in confidence judgements. *Acta Psychologica, 91*, 99-119.

Allwood, C.M., & Granhag, P.A. (1999). Feelings of confidence and the realism of confidence judgments in everyday life. In P. Juslin & H. Montgomery (Eds.), *Judgment and decision making: Neo-Brunswikian and process tracing approaches* (pp. 123-146). Hillsdale, N.J.: Lawrence Erlbaum Press.

Allwood, C.M., Granhag, P.A., & Johansson, M. (2003). Increased realism in eyewitness confidence judgments: The effect of dyadic collaboration. *Applied Cognitive Psychology, 17*, 545-561.

Andersson, J., & Rönnberg, J. (1997). Cued memory collaboration: Effects of friendship and type of retrieval cue. *European Journal of Cognitive Psychology, 9*, 273-287.

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

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121*, 149-167.

Granhag, P.A. (1997). Realism in eyewitness confidence as a function of type of event witnessed and repeated recall. *Journal of Applied Psychology, 82,* 599–613.

Granhag, P.A., Strömwall, L.A., & Allwood, C.M. (2000). Effects of reiteration, hindsight bias, and memory on realism in eyewitnesses' confidence. *Applied Cognitive Psychology, 14*, 397-420.

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.

Hinsz, V.B., Tindale, R.S., & Vollrath, D.A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin, 121*, 43-64.

Hertwig, R., Gigerenzer, G., & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychological Review, 104*, 194–202.

Johansson, M., Allwood, C.M., & Granhag, P.A. (2002). *Dyads discussing confidence: The effect of different types of working activities.* Unpublished manuscript, Department of psychology, Lund University and Department of psychology, Göteborg University, Sweden.

Kerr, N.L., MacCoun, R.J., & Kramer, G.P. (1996). "When are N heads better (or worse) than one?": Biased judgment in individuals versus groups. In E.H. Witte & J.H. Davies (Eds.), *Understanding group behavior: Consensual action by small groups* (Vol. 1, pp. 105-136). Mahwah, N.J.: Erlbaum.

Koehler, D.J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110*, 499-519.

Koehler, D.J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*, 461-469.

Lichtenstein, S., Fischhoff, B., & Phillips, L.D. (1982). Calibration of probabilities: The state of the art in 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306 – 334). Cambridge: Cambridge University Press.

McClelland, A.G.R., & Bolgar, F. (1994). The calibration of subjective probabilities: Theories and models 1980-1994. In G. Wright, & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). New York: John Wiley & Sons.

Sniezek, J., & Henry, R.A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43*, 1-28.

Stephenson, G.M., & Wagner, W. (1989). Origins of the misplaced confidence effect in collaborative recall. *Applied Cognitive Psychology, 3*, 227-236.

Treadwell, J.R., & Nelson, T.O. (1996). Availability of information and the aggregation of confidence in prior decisions. *Organizational Behavior and Human Decision Processes, 68*, 13-27.

Wilke, H., & Kaplan, M. (2001). Task creativity and social creativity in decision-making groups. In C.M. Allwood & Selart M. (Eds.), *Decision making: Social and creative dimensions* (pp. 35-51). Dordrecht: Kluwer Academic Publishers.

Yzerbyt, V.Y., Lories, G., & Dardenne, B. (Eds.). (1998). *Metacognition: Cognitive and social dimensions*. London: Sage.