AN ABSTRACT OF THE THESIS OF

Jackson Duncan-Reid for the degree of Master of Science in Psychology presented on June 11, 2019.

Title: Collaborative Metacognition and Decision Making in a Signal Detection Task

Abstract approved:

_____

Jason S. McCarley

When decision makers work collaboratively, their combined performance can exceed the performance of even the best group member working alone. However, despite the potential for performance gains when working in a group, collaborating individuals often show poor coordination and losses of motivation, and are generally inefficient at combining their resources and effort – all contributing to suboptimal group performance. The present experiment aimed to investigate how cognitive self-monitoring, or *metacognition*, changes under collaborative conditions. 54 participants (38 female, 16 male) formed 27, 2-person groups, and performed a gauge-aggregation signal detection task both individually and collaboratively. Measures of Type-1 (signal detection sensitivity) and Type-2 (metacognitive sensitivity) performance were calculated from individual and group confidence ratings. Hierarchical Bayesian parameter estimates suggested that when working collaboratively, groups showed no collaborative gains in task performance, but did trend towards outperforming the more metacognitively sensitive members. Similarly, groups showed no collaborative gains in metacognitive efficiency, and in fact trended toward metacognitive losses.

Despite the potential for collaborative gains and increased metacognitive awareness, groups likely failed to show any benefits of teamwork due to poor integration of information, and suboptimal weighting of group member contributions.

Collaborative Metacognition and Decision Making in a Signal Detection Task

by

Jackson Duncan-Reid

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented June 11, 2019
Commencement June 2019

Master of Science thesis of Jackson Duncan-Reid presented on June 11, 2019

APPROVED:

Major Professor, representing Psychology

Director of the School of Psychological Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Jackson Duncan-Reid, Author

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF APPENDICES

Collaborative Metacognition and Decision Making in a Signal Detection Task

When individuals work collaboratively, their combined performance has the potential to exceed what even the best group member could achieve on their own (Bahrami et al., 2012a; Sniezek & Henry, 1989; Sorkin, Hays, & West, 2001). Through the pooling of resources and effort, the sharing of unique information, and the utilization of group members' specific competencies, collaborative groups should, intuitively, show levels of performance above that of just a single person in any given task. Some evidence supports this assumption; the benefits of teamwork and collaboration have been observed across several task domains, with groups typically outperforming individuals in physical tasks (Kravitz & Martin, 1986), information recall (Weldon & Bellinger, 1997), vocabulary-based problem solving (Watson, 1928), and perceptual decision making (Bahrami et al., 2010; Sorkin et al., 2001). For example, when participants were instructed to recall as many items as possible from a previously observed sequence of photographs and names of common objects, groups of individuals working collaboratively recalled more unique items than the best individuals working alone (Weldon & Bellinger, 1997). In Sorkin et al.'s (2011) visual decision-making task, participants were required to aggregate multiple sources of probabilistic information represented by visual gauges, and judge whether they indicated a signal state (gauges showing values higher on average) or a noise state (gauges showing values lower on average). Compared to individuals, collaborating groups showed greater ability to correctly judge the state represented by the gauges, with group performance increasing as groups became larger (Sorkin et al., 2001).

However, despite the observed benefits of collaboration, groups typically fall short of ideal or optimal performance (Kerr & Tindale, 2004), and in some cases perform worse than what the best individual members could achieve alone (Bahrami et al., 2010). In the previously mentioned examples of studies showing collaborative benefits, while groups did outperform individuals, they fell short of optimal performance. In Weldon and Bellinger's (1997) free recall of common items task, for instance, collaborative groups performed worse than *nominal groups* (groups comprised of individuals working alone but their effort is artificially pooled together without collaboration) – that is, groups of individuals recalling and discussing previously observed items together performed worse than equally sized 'groups' of individuals who engaged in no collaboration, but had their unique recalled items combined. In the visual decision-making task of Sorkin et al. (2001), collaborative group performance fell short of ideal statistical predictions, and performed even worse than models which assume no discussion or integration and instead make a simple majority-rules decision from each individual's private decision. In these cases, though groups outperformed individuals, they performed more poorly than, in principle, they could have.

Contributors to Group Suboptimality

Research into group decision making has been extensive, with the literature examining group work and the influences of various physical and social environments, group composition of gender/race/age, impacts of power/dominance/leadership, expertise and experience, and individual personality traits (see Kerr and Tindale, 2004, and Levine and Moreland, 1990, for reviews). In

the group/team performance literature, what constitutes a ‘group’ can include any arrangement of two or more individuals, encompassing potentially hundreds of members. For the purposes of the present research, discussions of group collaboration will use the term ‘group’ to refer to collections of 10 or fewer individuals performing tasks while interacting. There is no clear definition of what constitutes a ‘small group’ (Shaw, 1980), but much of the research to be discussed has focused on groups of approximately this size, most likely due to practical limitations in recruiting and observing large numbers of participants. Specific to task performance in these small groups, past research has investigated some specific contributing factors towards the observed suboptimality of collaborating individuals:

Coordination and Productivity Loss

An early documented case of group inefficiency was reported by Ringlemann (1913), and later translated by Kravitz and Martin (1986). It compared the physical effort of individuals and groups when pulling ropes and pushing crossbars. In these tasks, the average physical effort exerted by groups was less than the average summed effort exerted by the same individuals in the groups when working alone. This suboptimality of groups was presumed to have been caused by ‘coordination loss’; it was possible that the individuals working together did not exert their effort at exactly the same time and in the exact same direction, leading to poor combination of individual effort. Steiner (1966), in developing models of group performance, described the discrepancy between potential group performance and actual group performance as ‘process loss’, which is in part comprised of this coordination loss. Similar process losses are seen in cognitive tasks. In studies of brainstorming and

idea-generation (Diehl & Stroebe, 1991; Paulus & Dzindolet, 1993), collaborating groups consistently generate fewer novel ideas and solutions to hypothetical problems than nominal groups; in the collaborating groups, coordination and productivity losses occurred as only a single member could talk at a time, thereby preventing the other members from both generating and presenting their own ideas while preoccupied with listening to the speaker. Group coordination losses in cognitive tasks can also be caused by inferior organization of interpersonal communication, inefficient pooling of resources, and poor allocation of subtasks to appropriate group members (Kerr & Tindale, 2004; Ivan D. Steiner, 1966).

Motivation Loss

In addition to requiring effective coordination of resources and effort, a group requires that its members are sufficiently motivated in order to perform tasks optimally. 'Social loafing' refers to the psychological phenomenon wherein individuals show reduced motivation and effort when working in groups with others compared to when working alone (Karau & Williams, 1993). While Ringlemann (1913) attributed the suboptimal physical effort of groups pulling ropes to coordination loss, reexaminations of the original findings have better explained this group suboptimality as being caused by motivation loss. In a study conducted by Ingham, Levinger, Graves, and Peckham (1974), groups of various sizes pulling a rope showed diminishing returns with each additional group member added, with group effort falling short of the expected sum of individual efforts. When individuals pulled the rope alone but blindfolded and believing themselves to be pulling alongside others, their individual effort was lower than when they believed they were

pulling alone – suggesting that simply believing to be working collaboratively diminishes individual efforts. Individuals exert less effort when in groups partly because they believe any lack of personal effort will be obfuscated by the effort of the other group members, and partly because any potential praise, recognition, or rewards for good performance is divided across the group, thereby diminishing personal incentives proportional to the size of the group (Harkins, 1987; Latané, Williams, & Harkins, 1979). The effects of social loafing on group performance extend to cognitive tasks, visual decision making, and even clinical judgments by therapists (Karau & Williams, 1993).

Weighting Inefficiency

When working as part of a team or group, any potential benefits of collaboration are dependent on the type of task performed, or the specific strategies employed; a single group might show different levels of performance depending on how a task is structured or approached. In many cognitive decision-making tasks, group members must evaluate and combine information to reach a single group decision. In Steiner's (1972) taxonomy of tasks, this form of decision-making task is termed *discretionary*, because group members can attempt the task in a number of different ways: Individual judgments can be combined *additively* such that each member's contribution to the final decision is weighted equally, the final decision can be *disjunctive* in that the group simply defers to the best member, or the group could reach a medium between the two – combining all members' contributions to make the final decision, but weighting those contributions proportional to the ability of each person (Shaw, 1980). If the weightings assigned to group members are poorly

calibrated to either the ability of the individual or the quality of their information, groups can fail to outperform the best member, and even may perform worse than the best member (Hertz, Romand-Monnier, Kyriakopoulou, & Bahrami, 2015; Mahmoodi et al., 2015). In research by Hertz et al. (2015) and Mahmoodi et al. (2015), individuals working in pairs performed visual search tasks, but the tasks were made more difficult for one of the individuals within each pair by displaying poorer visual information. An optimal weighting strategy would therefore weight the contribution of the individual with the poorer information less heavily in the final group decision. However, groups showed an overall *equality bias*, weighting the contributions of both members equally, thereby preventing any collaborative gains.

Signal Detection Theory

Within the literature surrounding small groups, a portion of research has used *signal detection theory* to compare the performance of individuals and groups (Hinsz, 1990; Sorkin & Dai, 1994; Sorkin et al., 2001). Signal detection theory is a widely-accepted method of assessing the ability of a person or instrument to discriminate between discrete states of the world on the basis of probabilistic evidence (Stanislaw & Todorov, 1999). The theory is most commonly used to measure decision makers' ability to distinguish two categories of event, labelled as 'signals' and 'noise', with signals being the stimulus to be detected, and noise being a different stimulus, background stimuli, or random sensory distortion. A signal detection theory analysis can be applied in any context wherein individuals must decide which of two alternatives is represented in a source of information that contains some uncertainty. An example of such a situation might be that of lie detection; a person describes

several events, and another individual must discriminate between truth (noise) and lies (signals). In a more high-stakes example, signal detection theory analyses have been applied to examine the ability of airport baggage screening operators to correctly discriminate between bags containing typical luggage (noise), and bags containing explosives (signals) (Wells & Bradley, 2012).

When evaluating the performance of an individual in these stimuli discrimination tasks, one metric of ability might be the proportion of correctly identified signals (e.g. the proportion of untruthful statements correctly identified as lies), and the correctly identified noise events (e.g. the truthful statements identified as truths). These values are called the *hit rate* and *correct rejection rate*, respectively. The problem with this approach is that it does not account for *response bias* - the tendency for an individual to consistently respond in a certain way. If two individuals completed this hypothetical lie-detection task, one might show a higher hit rate than the other, suggesting a greater ability to detect lies, when in fact both individuals were equally good at discriminating truth from lies. The difference in their hit rates would lie in their response bias, with one individual being more distrustful and therefore more likely to say any given statement was a lie. The tendency to respond signal more than noise (or vice versa) can be influenced by the relative frequency of signal and noise events, termed the *base rate*, or by the consequences of making an incorrect decision, termed the *payoff*. The prevalence of airport baggage that contains explosives is very low, so to avoid wasting time and resources manually searching every bag that contains a potential explosive, one operator might take this base rate into consideration and be more likely to classify baggage as just showing noise.

Conversely, another operator might decide that a missed actual explosive could result in multiple deaths, and so takes this payoff into consideration and tends towards classifying bags with any uncertainty as containing a signal. Without a method of analysis that can separate discriminative ability from an individual's response bias, it is difficult and sometimes impossible to discern if two individuals show different hit rates because of different sensitivities, or different biases.

Signal detection theory offers a method of measurement that isolates the ability to discriminate between stimuli, called *sensitivity*, from response bias. In a signal detection task, the *decision variable* is the totality of evidence extracted from the information presented; in discriminating between lies and truth, the decision variable would be the subjective feeling of 'untruthfulness' present in a statement (which could be discerned from objective features like averted gaze, idle fidgeting etc.). In an objective case, the decision variable could be the number of antibody nanograms per milliliter in a substance-use test. The standard signal detection theory model assumes that signals and noise are normally distributed and have equal variance, and this model gives the following measures: a sensitivity index measure, d' (d prime), is a measure of the distance in standard deviation units between the peak of the normal distribution of noise values and the peak of the normal distribution of signal values. The measure d' is calculated by subtracting the z-scored hit rate from the z-scored *false alarm rate* (the proportion of noise incorrectly labelled as signals). Bias is measured by the *criterion*, termed *c*, which the individual's cutoff point of a decision variable where any value greater than this criterion is classified as a signal. Together, d' represents the ability to separate two stimuli, and *c* represents the degree

of bias towards responding signal or noise. A visual representation of signal detection theory measures is presented in Figure 1.

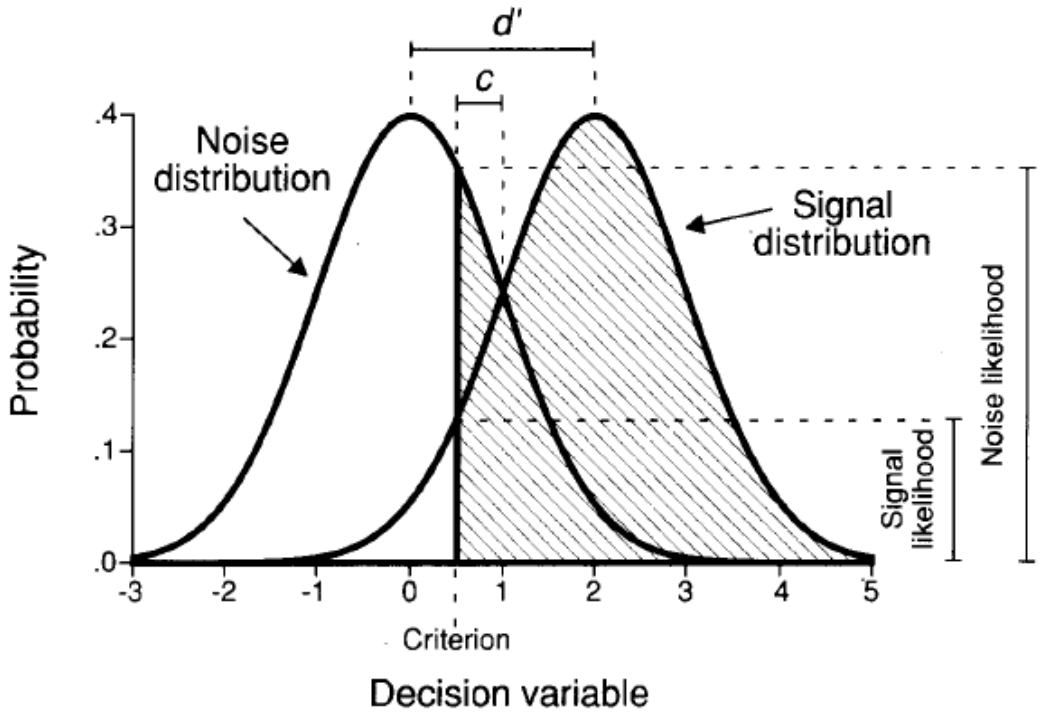

Figure 1. Signal detection theory measures: *d′* (sensitivity) and *c* (bias), (from Stanislaw & Todorov, 1999).

When applied in the context of group performance, group *d′* represents the ability of the group members to integrate their separate sources of information, and successfully separate signals from noise.

Metacognition

When individuals perform cognitive tasks both on their own and within groups, they not only engage in coordination of their effort and resources, but they also monitor their own cognitions and performance. Cognitive self-monitoring, or *metacognition*, involves the observation, reflection, and regulation of one's own

thought processes, knowledge, and cognitive strategies (Flavell, 1979). Metacognition is of interest to researchers because of its relationship with task performance; groups with greater metacognitive knowledge show greater performance in spatial decision making tasks (Hamilton, Mancuso, Mohammed, Tesler, & McNeese, 2017), individuals with poor performance and skill show correspondingly poor metacognitive awareness of their own abilities (Kruger & Dunning, 1999) and greater metacognition allows for better adaptability and use of cognitive strategies (Flavell, 1979). For instance, in a study of metacognitive awareness and task strategy selection, participants performed a flanker task in two separate but seemingly identical configurations, and in each trial block could choose which task configuration they would prefer to complete (Desender, Buc Calderon, Van Opstal, & Van den Bussche, 2017). Participants who were classified by independent raters as 'metacognitively aware' (compared to unaware participants) were able to recognize that one configuration was covertly more difficult than the other, and consistently chose the easier configuration – leading to greater task performance, while metacognitively unaware participants noticed no differences in the configurations, chose them arbitrarily, and performed worse as a result.

What encompasses 'metacognition' is widely debated, however. Multiple definitions of the term exist (Beran, Perner, & Proust, 2012; Garofalo & Lester, 1985), and some convincing arguments have been put forward that at least one aspect of metacognition, introspection about the reasoning behind our own behavior, is misleading at best, if not impossible in humans (Nisbett & Wilson, 1977). For the purposes of the present study, metacognition is conceptualized as an individual's

knowledge and awareness of their own performance and ability within the context of a signal detection decision-making task.

Past research into group metacognition has typically examined self-reported, subjective confidence judgments. Bahrami et al. (2012b) examined the role of metacognition in collective decision making; groups of individuals were assigned different methods of communication, and correlations between metacognitive alignment within the groups and performance suggested that groups sharing metacognitive information non-verbally through confidence ratings achieved greater signal detection performance. Hinsz (1990) explored the role of metacognition in group recognition memory, with gamma rank correlations suggesting group confidence was better calibrated to memory recall performance than was individual confidence. A study by Sniezek and Henry (1989) required individuals and groups to estimate unknown quantities; participants constructed 99% confidence intervals around their best estimate of the number of deaths per year in the United States from assorted causes (e.g. homicide, heart disease). While groups were more accurate in their final estimates and more confident in their decisions than were individuals, neither the confidence of groups nor individuals was significantly correlated with accuracy – suggesting that only performance, not metacognition, showed a collaborative benefit.

However, correlations between confidence and task performance are poor measures of metacognitive ability. Nelson (1984), in a review of 'feelings of knowing' measures, argued that approaches using Pearson *r* correlations between confidence and accuracy to represent metacognition are potentially confounded by

response bias; differences in correlations between conditions/subjects may be due to disparate metacognition, or may be instead due to differing confidence thresholds. Nelson (1984) instead endorsed the gamma correlation measure as the recommended method of assessing feelings of knowing, citing in part its ease of interpretation and ability to be calculated from ordinal variables. Maniscalco and Lau (2012) however, suggested that gamma rank correlations between confidence and performance are also methodologically flawed, and do not separate confidence accuracy from changes in confidence criterions.

To address these issues with correlational measures, Maniscalco and Lau (2012) proposed a novel signal detection theory measure, *meta-d′*, which isolates metacognitive sensitivity from response bias in the same manner that *d′* isolates sensitivity from bias. When individuals perform a signal detection task, they make their signal/noise judgments (Type 1 sensitivity), and subsequently make a confidence judgment of the decision they just made (Type 2 sensitivity). This confidence judgment can be conceptualized as another signal detection judgment. However, rather than discriminating between signals and noise events in the world, individuals are now discriminating between their own correct and incorrect decisions. Metacognitive efficiency is then calculated by observing the discrepancy between the Type 1 and Type 2 sensitivities. If an individual's *meta-d′* is equal to their *d′* (*meta-d′/ d′* = 1), they are considered metacognitively ideal. If the ratio of *meta-d'* to *d'* is less than 1, this represent the relative metacognitive inefficiency of the individual. *Meta-d'* can be lower than *d'* if noise is introduced in the confidence rating process; in a study investigating the effects of fatigue on perceptual sensitivity and metacognitive

vigilance in a signal detection task, *meta-d′* decreased over time at greater rate than did *d′*, suggesting that fatigue introduces noise into both the processes of perception and metacognition, but metacognitive processes are affected more harshly and decline earlier (Maniscalco, McCurdy, Odegaard, & Lau, 2017). Differences in goal setting might also change *meta-d′* but not *d′*, as Maniscalco et al. (2017) argued that the fatigue-decline in metacognition could also have been explained by participants recognizing their fatigue, and deciding to prioritize their dwindling cognitive resources on the initial perceptual judgment, thereby neglecting metacognitive vigilance in favor of greater task performance.

Research has begun to use *meta-d'* and the metacognitive efficiency ratio within the signal detection framework to study metacognitive awareness. One such study by Palmer, David, and Fleming (2014) investigated metacognitive efficiency in a visual perception task with older and younger adults. Their design adjusted difficulty trial-by-trial to ensure that all participants had equivalent accuracy regardless of age, and compared metacognitive efficiency across the lifespan, finding that older adults show age-related declines of metacognitive efficiency irrespective of Type 1 sensitivity. *Meta-d'* has also been used to examine associations between arousal and confidence (Allen et al., 2016), and the relationship between metacognition, control of attention, and awareness of mind wandering (Desender et al., 2017).

Overall, metacognitive efficiency is a versatile measure of metacognitive ability that addresses the problems with the previous, potentially confounded correlational research, and is intuitive to understand and interpret.

The Present Study

The present study aimed to investigate if collaborating groups achieve metacognitive awareness above and beyond that of individuals working alone. Past research into the comparative metacognition of individuals and collaborating groups has utilized correlational measures that cannot separate actual metacognitive ability from the tendency to respond with bias in confidence-rating measures. Using Maniscalco and Lau's (2012) proposed metacognitive efficiency measure that isolates metacognitive sensitivity from response bias, the present study aimed to replicate and extend on a preliminary study investigating individual and group performance and metacognition in a signal detection task (Duncan-Reid & McCarley, 2017). The methods and analyses employed in this study were preregistered prior to data collection, with the registration publicly available from the Open Science Framework: https://osf.io/jsfwx

Hypotheses.

Hypothesis 1: Group *d'* will be higher than individual *d'*.

Given that the past research into group performance in signal detection tasks has reliably found that groups outperform individuals in Type 1 sensitivity, it was predicted that a similar pattern of collaborative gain would be observed in this study.

Hypothesis 2: *Meta-d'* will not differ credibly between the metacognitively better group members and the groups.

The preliminary research into group metacognition that the present study seeks to replicate has suggested that groups may engage in a strategy of metacognitive deferral, allowing the more metacognitively sensitive member to make

the judgments of group confidence, thereby achieving no collaborative gains in metacognition (Duncan-Reid & McCarley, 2017).

Hypothesis 3: *Meta-d'* of the metacognitively worse group members will be lower than that of the groups.

As analyses compared the metacognitively worse members and metacognitively better members to the groups separately, it was predicted that groups would show metacognitive sensitivity greater than that of the worse members.

Hypothesis 4: Group metacognitive efficiency will be greater than the metacognitive efficiency of the worse member, but not credibly different from that of the better member.

As groups have been found to be proficient in correctly identifying their most capable members (Henry, 1993), it was predicted that groups would show a strategy of deferral to the better member, outperforming the worst members but not exceeding the metacognitive ability of the better members.

Methods

Participants

Participants were 54 introductory psychology students that formed 27 groups (38 females and 16 males, mean age = 19.24, SD = 1.25, range = 18 - 23 years old) recruited through the School of Psychology online experiment sign-up system, and were compensated with partial course credit for undergraduate psychology courses. Participants were screened for normal or corrected-to-normal visual acuity with a Snellen letter chart, and normal colour vision with Ishihara colour-deficiency test plates. Data from an additional 11 groups (22 participants total) were not included in

the final sample of 54 participants, due to one or more of the group members failing to meet the vision requirements.

Apparatus

The experimental stimuli were presented on 24” ViewSonic XG2401 LED color monitors driven by Apple Mac mini desktop computers. Monitors were set to 1920x1080 resolution at a 16:9 aspect ratio, with a 60-Hz refresh rate. Viewing distance from the screen was approximately 65 cm. The experiment was coded and performed in the PsychoPy software package (Peirce, 2007).

Each stimulus display was comprised of five vertical, rectangular gauges presented on a dark grey (#3f3f3f) background. Gauges were 82 x 13 mm in size, drawn in white (#ffffff) 2-pixel stroke, and subdivided into 10 equal regions by 8 white horizontal tick marks. Each gauge included a red (#ff0000) horizontal line segment, 15 in length and drawn in 1-pixel stroke, that served as a moving marker.

Procedure

Participants performed a multi-cue, two-alternative yes/no signal detection task, adopted from Montgomery and Sorkin (1996). Participants were instructed to imagine they were working in a nuclear power plant and were required to judge each trial whether the state of the system was normal (noise) or dangerous (noise + signal). Participants were presented with 5 analog gauges and were tasked with aggregating the individual gauge readings to judge whether they indicated a noise state (readings were on average below the middle line) or signal state (readings were on average above the middle line). This task was probabilistic in nature; while the true mean of the signal distribution lay above the middle line, it is possible that the average reading

on a given signal trial is below the line, requiring participants to account for this uncertainty and preventing perfect performance on the task. Figure 2 presents two sets of example stimuli. On noise trials, the readings of the three leftmost gauges were independent and identically distributed (i.i.d) with value $X \sim N(-10, 100)$, and the readings of the two rightmost gauges were i.i.d with values $X \sim N(-10, 400)$. On signal trials, the readings of the three leftmost gauges were i.i.d with value $X \sim N(10, 100)$, and the readings of the two rightmost gauges were i.i.d with values $X \sim N(10, 400)$. The three leftmost gauges and the two rightmost gauges were assigned different variances to add a moderate level of task difficulty; in order to reach an optimal decision, participants needed to differentially weight the separate gauges when aggregating the readings. The distribution of evidence values across the five gauges allowed a maximum sensitivity of $d'_{ideal} = 3.74$. When working in a group, participants were each shown an independent set of 5 gauges displaying readings drawn from the same state distribution, and were required to make their group judgment based on the pattern of values across all 10 gauges. The distribution of evidence values across the ten gauges allowed a maximum group sensitivity of $d'_{ideal} = 5.29$.

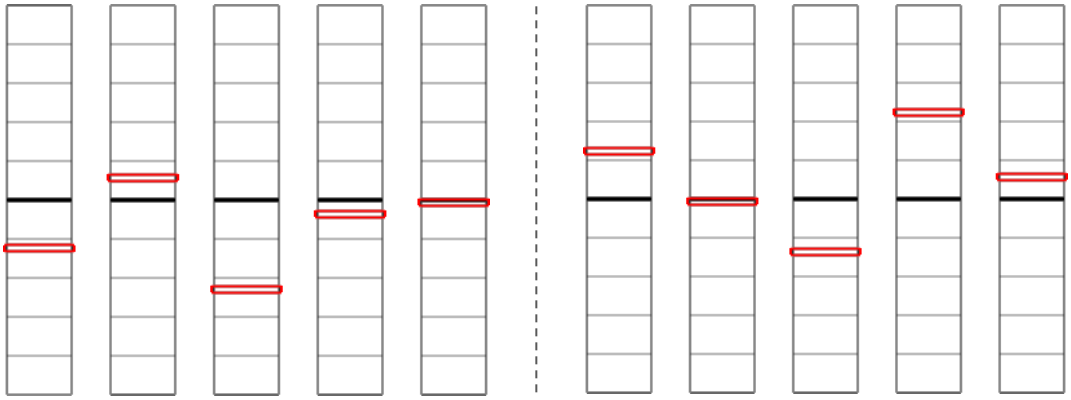

Figure 2. Example sets of gauge stimuli from the experiment showing a noise trial (left group of 5 gauges), and a signal trial (right group of 5 gauges), black and white colours are inverted for clarity.

Instructions to participants explained that on average, gauge values (indicated by the red horizontal markers) would be lower than the midpoints of the gauges if the state of the system was normal and above the midpoints if the state of the system was dangerous, and that the two rightmost gauges were less reliable than the three leftmost gauges.

Participants made their judgments by mouse-clicking on one of two buttons, labelled ‘normal’ or ‘dangerous’, located directly below the gauges. They were subsequently prompted to rate their response confidence on a 6-point scale, with endpoints labelled ‘Not at all Confident’ and ‘Very Confident’. The red horizontal markers were removed from the screen immediately following the initial binary response, to discourage participants from collecting further information from the display after they had made their binary judgment. A feedback message immediately followed the confidence rating. The feedback message read ‘Good Judgment!’

following a hit or correct rejection, “Oops! The situation was normal.” following a false alarm, and “Oops! Readings were too high!” following a miss.

After arriving at the lab participants, read and signed a consent form, then performed the visual acuity and colour vision screening tests. When both participants were present, they were introduced to one another by name and instructed to sit at separate computers, side-by-side on the same table. A corkboard separating the monitors ensured that although participants could see one another, neither could see the others’ display. Instructions were presented to the subjects on PowerPoint slides. The participants were asked to read through the instructions individually, each clicking through the slides at their own pace. After reading the instructions, participants performed a 5-minute practice block of trials individually, followed by a 5-minute practice block of trials working collaboratively. When working individually, participants were instructed, “Please don’t talk or discuss the task in any way with the other participant,” and when working in a group they were instructed to, “Discuss the task verbally however [they] like.” During the collaborative blocks, the group member sitting at the rightmost computer controlled the mouse and entered the group’s judgment and confidence rating. They were instructed that “Whoever is sitting at the computer on the right has control of the mouse, but isn’t making the final decision, just clicking the buttons on behalf of the group.” A diagram of the experimental setup is presented in Figure 3.

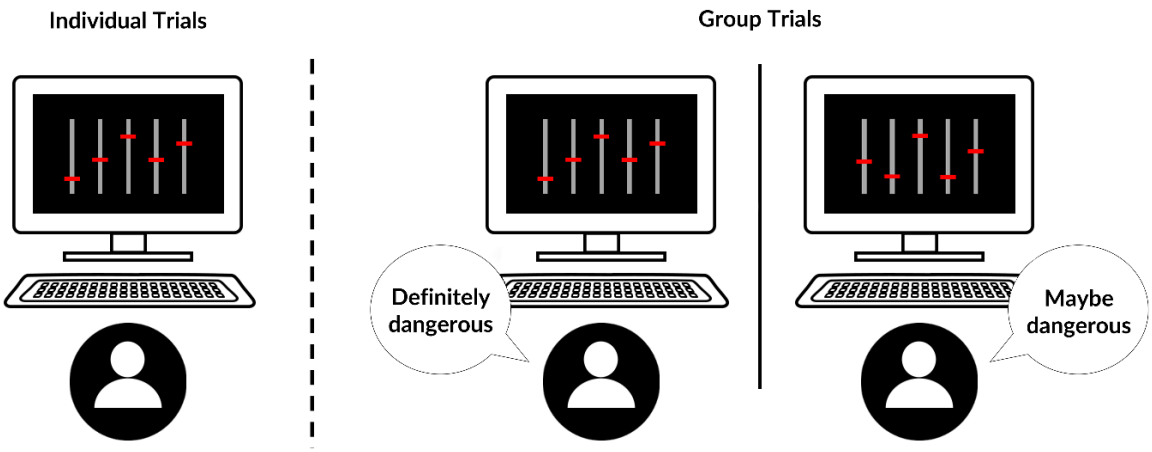

Figure 3. A diagram representing a participant working independently in an individual trial (left), and participants verbally discussing their unique gauges in a group trial (right).

After the two practice blocks of trials, each pair of participants completed two experimental blocks of the task individually for 6 minutes and 30 seconds each, and two experimental blocks working in a group for 8 minutes and 30 seconds each. The extra time allotted for the group blocks aimed to ensure a similar number of trials in each condition, accounting for the extra time required for the group members to discuss their judgments and make a joint decision. The order of the group and individual experimental blocks was determined by an ABBA design, randomly counterbalanced across the participants. Within the second experimental block of group trials, participants were instructed to swap computers so that the leftmost participant would then control the mouse to indicate the final decision on behalf of the group.

Statistical Analyses

The following analyses were conducted with a hierarchical Bayesian parameter estimation procedure (Kruschke, 2013). When performing Bayesian estimation, an initial distribution of parameter values is assumed; this is called the *prior distribution*. A Markov Chain Monte Carlo sampling procedure subsequently reallocates credibility, preferring parameter values that are most likely given the data. The resulting parameter estimates form the *posterior distribution*. Here, prior distributions were intentionally vague, in order to allow the data to overwhelm this initial allocation of credibility

The data reported in the present study are the posterior distributions of the hierarchically estimated grand means, with Bayesian credible intervals. The Type 1 sensitivity measure *d'*, and the Type 2 sensitivity measure *meta-d'* proposed by Maniscalco & Lau (2012) were each calculated separately for each individual and group using Fleming's (2017) Bayesian estimation procedures. These measures were subsequently analyzed in a hierarchical one-way design with three levels, using the model recommended by Kruschke (2014).

Results

A common benchmark employed in small group research to determine if groups show a collaborative gain is to examine whether the group achieves performance above that of the best group member. As metacognition was the main focus of the present research, in order to determine if a group produced a metacognitive gain above that of the best member in the group, analyses ranked the members of each group according to their *meta-d'*, and compared group performance

to the better and worse members individually. The means and corresponding Bayesian credible intervals for the following analyses are reported in Table A.1 of the Appendix.

Figure 4 presents posterior distributions of *d′* for the metacognitively worse members, better members, and the groups. Figure 5 presents posterior distributions for the *difference* in *d′* between the groups, and the metacognitively better and worse members. Hypothesis 1 predicted that group *d′* would be greater than that of the individuals. This prediction was only partially supported; the posterior distribution of the difference between the *d′* of the metacognitively worse members and the groups showed a credible difference (M_{diff} = 0.53, BCI = [0.28, 0.77]) with groups showing greater Type 1 sensitivity than the metacognitively worse members. However, despite the data indicating a trend towards the groups displaying a collaborative gain, there was no credible difference observed between the *d′* of the better members and the groups, as the credible interval overlapped 0 (M_{diff} = 0.19, BCI = [-0.04, 0.42]). The metacognitively better members showed greater Type 1 sensitivity than the metacognitively worse members (M_{diff} = 0.34, BCI = [0.10, 0.57]).

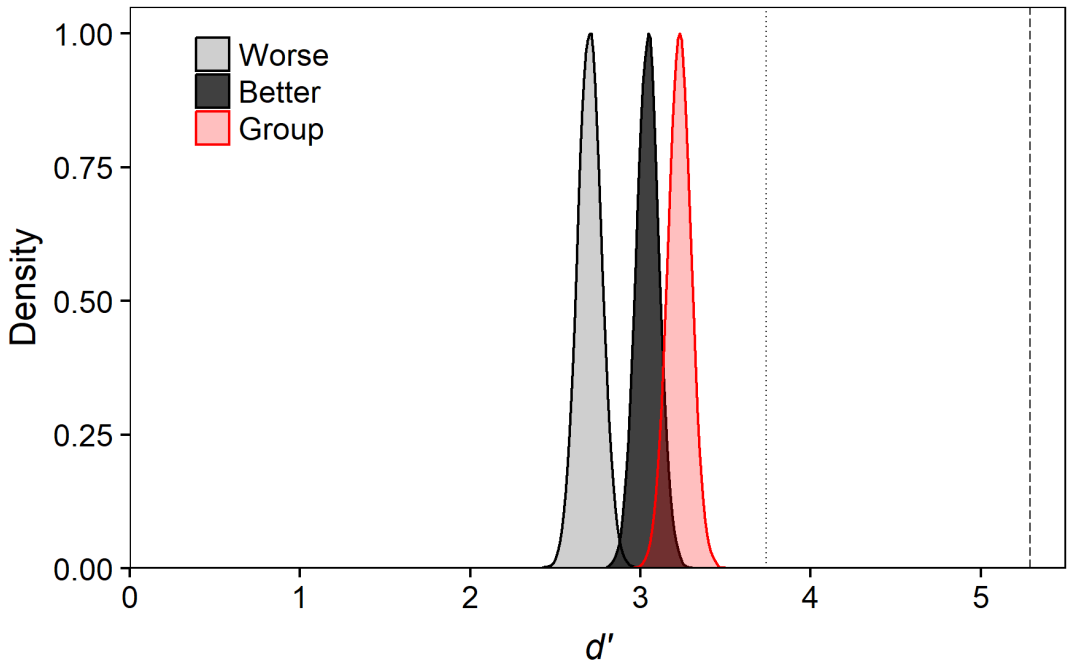

Figure 4. Posterior distributions of *d'* (sensitivity) for the metacognitively worse individuals, better individuals, and groups. The leftmost dotted line indicates the maximum achievable *d'* for the individuals, and the rightmost dashed line indicates the maximum achievable *d'* for the groups.

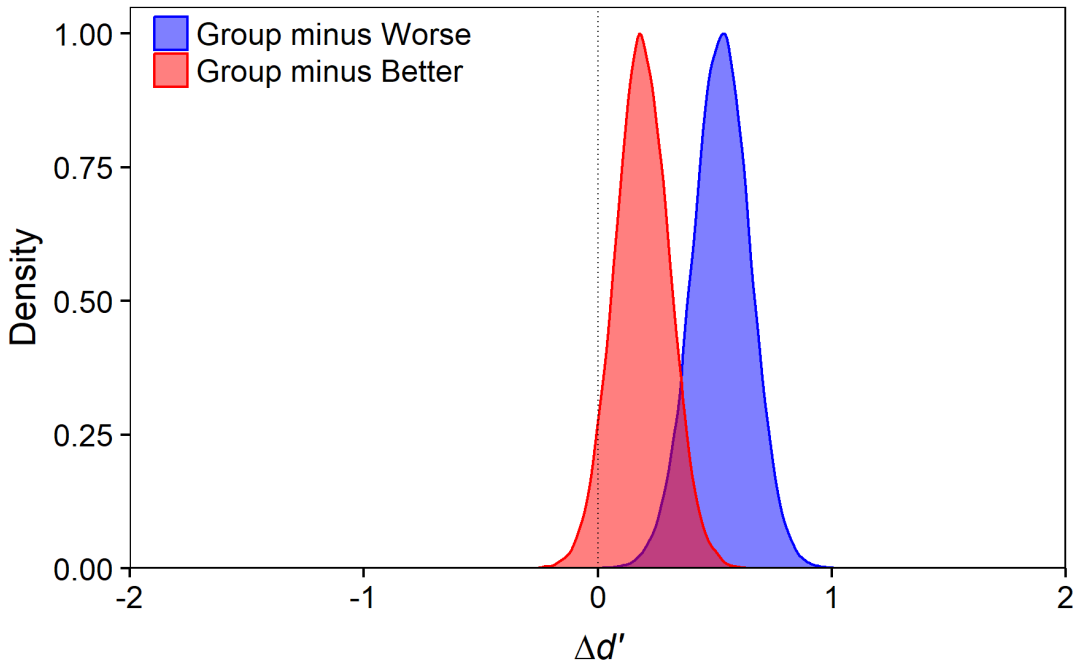

Figure 5. Posterior distributions of *d'* differences between the groups, and both better and worse members. The dotted vertical line indicates a mean difference of 0.

Figure 6 presents posterior distributions of *meta-d'* for the metacognitively worse members, better members, and the groups. Figure 7 presents posterior distributions for the *difference* in *meta-d'* between the groups, and the metacognitively better and worse members. Hypothesis 2 predicted that *meta-d'* would not differ credibly between the metacognitively better members and the groups. This hypothesis was supported; the posterior distribution of the difference between the *meta-d'* of the metacognitively better members and the groups showed no credible difference (M_{diff} = -0.10, BCI = [-0.54, 0.33]). In this experiment however, Bayes factors or similar methods of assessing null values were not calculated or planned ahead of data collection (e.g. Regions of practical equivalence, Savage-

Dickey Bayes factor approximations), so it is uncertain wether there is no difference between the *meta-d′* of the better members and the groups. Hypothesis 3 predicted that the *meta-d′* of the metacognitively worse group members would be lower than that of the group. This hypothesis was supported; groups showed greater Type 2 sensitivity than the meta-cognitively worse members (M_{diff} = 0.70, BCI = [0.24, 1.16]). Unsurprisingly, a credible difference in *meta-d′* was observed between the metacognitively worse members and the metacognitively better members, (M_{diff} = 0.80, BCI = [0.33, 1.27]), with the better members outperforming the worse members in Type 2 sensitivity.

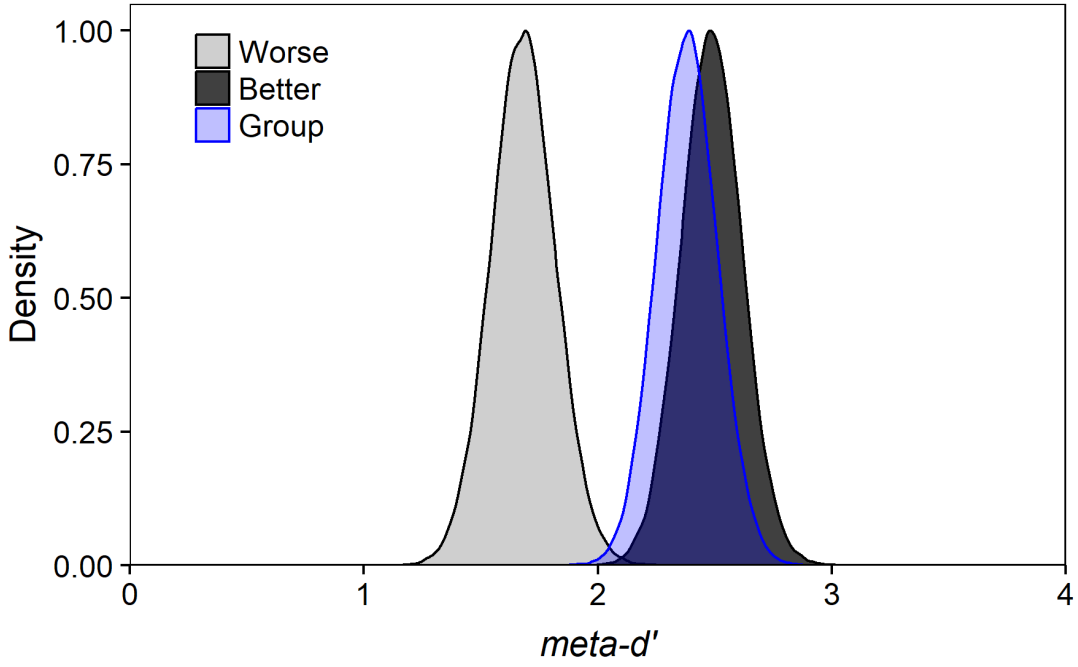

Figure 6. Posterior distributions of *meta-d′* (metacognitive sensitivity) for the metacognitively worse individuals, better individuals, and groups.

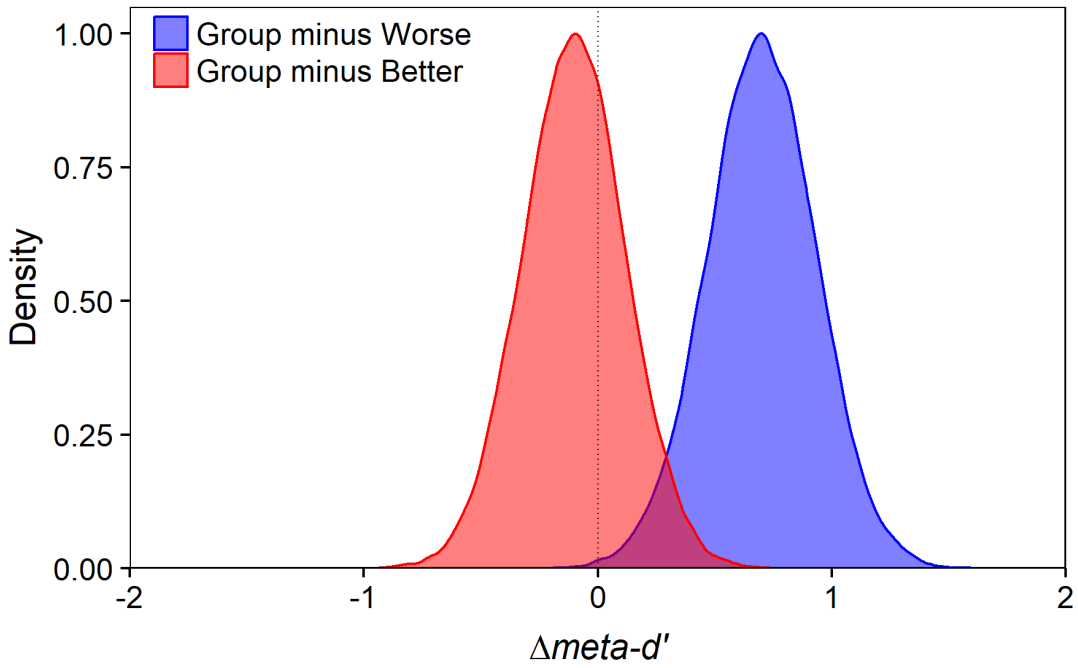

Figure 7. Posterior distributions of *meta-d'* differences between the groups, and both better and worse members. The dotted vertical line indicates a mean difference of 0.

Figure 8 presents posterior distributions of metacognitive efficiency for the metacognitively worse members, better members, and groups. Figure 9 presents posterior distributions for the *difference* in metacognitive efficiency between the groups, and the metacognitively better and worse members. Hypothesis 4 predicted that groups would show greater metacognitive efficiency than the metacognitively worse group members but would be not be credibly different from the metacognitively better group members. This hypothesis was only partially supported. Despite the data indicating a trend towards the groups outperforming the metacognitively worse members in metacognitive efficiency, the posterior distribution of the difference between the worse members and the groups showed no

credible difference (M_{diff} = 0.12, BCI = [-0.05, 0.28]). As was predicted, there was no credible difference between the metacognitively better members and the groups (M_{diff} = -.08, BCI = [-0.23, 0.07]), however, the data suggest a trend towards the groups showing *worse* metacognitive efficiency than the better members. Unsurprisingly a credible difference in metacognitive efficiency was observed between the metacognitively better members and the worse members, (M_{diff} = 0.19, BCI = [0.02, 0.37]), with the metacognitively better members showing greater metacognitive efficiency than the worse members.

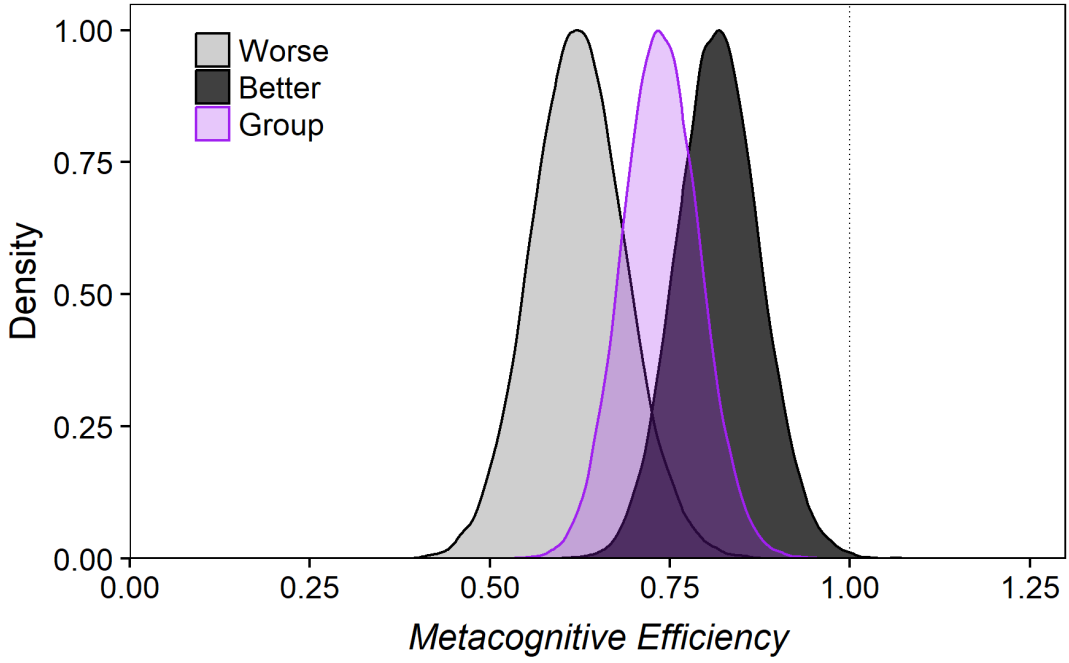

Figure 8. Posterior distributions of metacognitive efficiency (*meta-d'/d'*) for the metacognitively worse individuals, better individuals, and groups. The dotted vertical line indicates ideal metacognitive efficiency of 1.

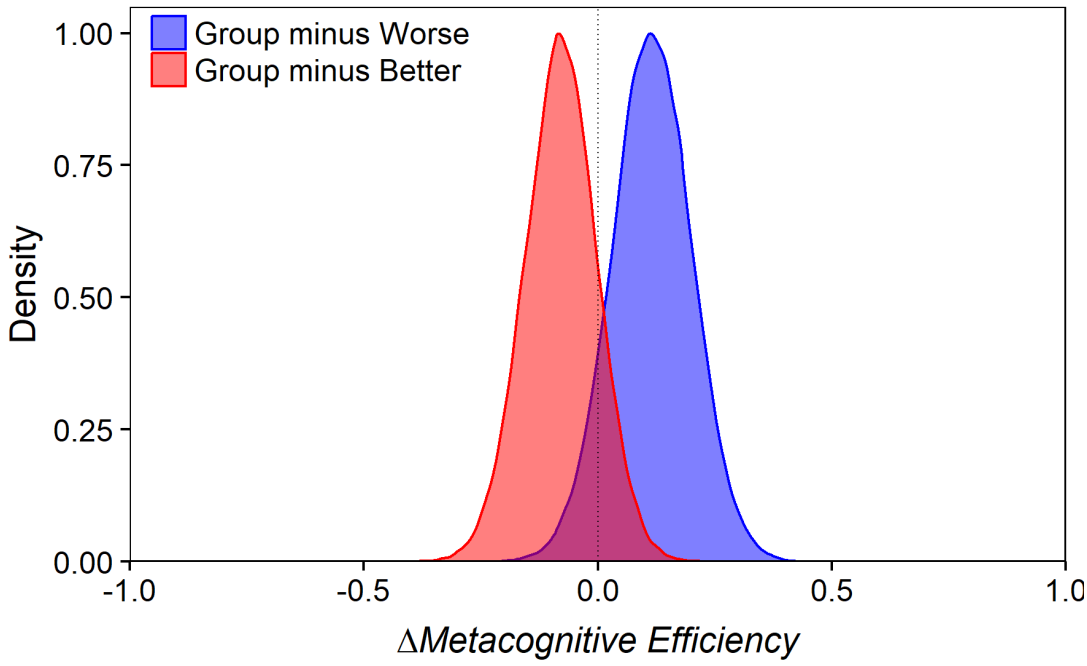

Figure 9. Posterior distributions of metacognitive efficiency differences between the groups, and both better and worse members. The dotted vertical line indicates a mean difference of 0.

Discussion

When individuals work collaboratively in cognitive tasks, they share, discuss, and integrate information to inform their decisions as a group. As part of this collaborative process, group members also monitor their own performance and cognitions to inform their behaviour and strategy. Although research has shown that groups can achieve collaborative gains in performance, it is unclear whether they demonstrate collaborative gains in metacognitive ability. Past research that has examined group metacognition has attempted to measure it through correlations between confidence and performance, an approach that is methodologically flawed, conflating metacognitive ability with response bias. The present study compared

individual and group sensitivity, and metacognitive efficiency within a signal detection theory framework that addresses the potential confounds of past research.

Groups outperformed the metacognitively worse members in Type 1 sensitivity, but surprisingly did not show a collaborative benefit above the performance of the better members, however there was a trend towards outperforming the better members. Type 2 metacognitive sensitivity also showed a similar pattern; groups outperformed the worse members, but was no different than that of the better members. Group metacognitive efficiency was not greater than the best members, and groups instead trended towards being less metacognitively efficient than the best members.

Sensitivity

Contrary to what was predicted in Hypothesis 1, groups did not show a collaborative benefit in Type 1 sensitivity. This is surprising, as research into group performance in signal detection tasks has typically found that groups show greater sensitivity than individuals (Bahrami et al., 2010; Sorkin et al., 2001). Similarly, the past research this experiment is attempting to replicate did find a collaborative benefit above that of the better members. The information available to the groups was much greater than individuals working alone (groups were presented with 10 gauges total, individuals only 5 gauges), so groups should ideally have shown greater performance than the best individuals. While groups did show a trend towards outperforming the better members, they still fell drastically short of ideal performance, with the average group d' (M = 3.23, BCI = [3.06, 3.39]) less than the maximum achievable group d' (d'_{ideal} = 5.29).

One explanation for these results is that the groups were inefficient due to poor collaborative strategies. When discussing the information given by the separate gauge displays, participants may have used inferior verbal communication that contributed to losses in information that contributed to the final group decision. They may also have assigned suboptimal weights to the contributions of each group member; the best performing individual within the group (and the information they extract from the display) should ideally have more weight in the final group decision than the poorer performing member. In the signal detection task of Bahrami et al. (2012b), the poorer performing member of 2-person groups had additional noise introduced into their display during collaboration to artificially widen the performance gap between individuals. Verbally communicating, but not non-verbally communicating groups, performed worse than the better members, suggesting that specifically the verbal discussion that facilitates collaboration contributes to inefficient weighting of member contributions to the final decision. Whether participants in the present study exhibited equality bias and implicitly assumed they were of equal competence (Bahrami et al., 2012b), or recognized differences in ability but failed to account for them in an effort to avoid the uncomfortable social exclusion elicited by overruling or ignoring the inferior group member's decisions (Mahmoodi et al., 2015), the groups may have used a suboptimal method of weighting that prevented performance gains.

Another possible explanation for this lack of collaborative gain is that participants did not have enough experience with the signal detection task itself, or with working with their fellow group member, to collaborate effectively and combine

their judgments in an ideal manner. The signal detection task used by Sorkin et al. (2001) that found collaborative gains in 2-person groups shared some similarity to the present study; individuals and groups aggregated readings from 9 analog gauges displayed for 320ms. In Sorkin et al.'s (2001) study however, participants completed 12,000 individual trials of the signal detection task across 6 sessions that took approximately 1.5 hours each, then completed 800 trials working as groups. In the present study participants completed 2, 5-minute blocks of practice trials, and then completed counterbalanced individual and group blocks for 30 minutes, averaging 171 trials working as individuals and 108 trials working as groups. While the design of the Sorkin et al. (2001) experiment might not be directly comparable to the present study, each member had completed thousands of trials and many more hours of experience with the task by the time they collaborated with another participant, and collaborated for several hundred more trials as groups than the present study. This additional time performing the task may have allowed participants to develop a greater understanding of the task and how to best perform it, as well as granted more experience working with their other group member to develop effective coordination strategies that allowed for a collaborative benefit to be achieved.

Bahrami et al. (2012a) demonstrated collaborative performance gains over time in 2-person groups completing a Gabor patches signal detection task; the results of this experiment suggested that when group members collaborated and were presented trial-by-trial feedback, they showed collaborative gains quickly from the beginning of the task, and gains increased until plateauing after approximately 130 trials. This suggests benefits of collaboration can and should be observed early after

only a short time of working with another person. It is therefore surprising that in the present study, groups only showed a trend towards a collaborative gain and were not credibly different than the better members, even with 22 minutes of collaboration. Again, the particular signal detection task used by Bahrami et al. (2012a) might not be completely comparable to the present experiment, as performance on a Gabor patches task depends solely on extracting visual information from a stimulus presented for a short time. The gauges task of the present experiment required the extraction of visual information, aggregation of readings, weighting of reliable/unreliable gauges, and finally the integration of information communicated by another individual in order to make an accurate decision. This demanding process might require greater practice and time spent working with a fellow group member to develop an effective system of coordination than other types of collaborative tasks. Research that previously demonstrated collaborative gains involve simple tasks that come naturally to participants and require little previous experience; pulling ropes, remembering words, and brainstorming ideas for example. The present study is more cognitively demanding than these common tasks, and aggregating/weighting gauges to judge the state of a nuclear power plant is likely to be a task unfamiliar to most participants and requires more practice and experience to perform ideally.

Beyond the design and cognitive requirements of the gauges task itself, losses of motivation may also have contributed to the suboptimality of groups. Members of groups have been suggested to engage in social loafing not only because their individual effort cannot be directly observed in some group situations (allowing them to exert less effort without consequences), but because of submaximal goal setting

(Latané et al., 1979). Rather than trying to perform at their absolute best possible level as a group, participants may have simply aimed to perform ‘well enough’. With twice the number of gauges presented to the groups than presented to individuals, aggregation and weighting of the information can be performed less diligently by the group and still reach performance comparable to an individual. It is possible that participants may aim to achieve an arbitrary level of performance they deem acceptable, and rather than exceed that level when collaborating now that more information is available, they decrease their effort as this level of performance is achieved more easily. As performing the gauges task in the present study is cognitively demanding (and arguably monotonous to complete for hundreds of trials), unpaid, unrewarded participants with no consequences for poor performance may become fatigued after several blocks of trials and aim to optimize their effort for an acceptable standard rather than aim to maximize their group performance.

Metacognition

As was predicted by Hypotheses 2 and 3, the *meta-d′* of the groups exceeded the *meta-d′* of the metacognitively worse individuals, but was no different than the *meta-d′* of the better members. *Meta-d′* represents the ability to discriminate between one’s own correct and incorrect decisions and is compared against *d′* to calculate metacognitive efficiency - a measure indicating the efficiency of an individual’s metacognition as a ratio. *Meta-d′* loses interpretability when not evaluated in this manner relative to *d′,* so the focus of this study concerns metacognitive efficiency. Hypothesis 4 was only partially supported; the metacognitive efficiency of groups did not exceed that of the metacognitively worse members, and group metacognitive

efficiency was no different than the metacognitive better members, instead showing a trend towards groups being metacognitively worse than the better members. It is important to reiterate that since the signal detection theory measures for Type 1 sensitivity and Type 2 metacognitive sensitivity isolate sensitivity from response bias, any differences across conditions cannot be attributed to differences in bias. Therefore, a change in metacognitive awareness cannot be explained by a confidence shift in a singular direction (e.g. an individual who becomes more confident when working in a group). If 'Individual A' demonstrates lower metacognitive awareness after some experimental manipulation, this lower metacognitive awareness is caused by their becoming systematically overconfident when they were incorrect in their Type 1 decision, and systematically underconfident when they were correct.

In the present study, groups did not show a collaborative gain in metacognitive efficiency above that of the better members. This lack of a metacognitive gain could be attributed to the strategy used by the groups to reach a consensus regarding their confidence on any given trial. The preliminary research into the metacognitive efficiency of groups that this study sought to replicate and extend suggested that groups used a strategy of metacognitive deferral; over the course of the experiment, through communicating their confidence and receiving trial-by-trial feedback, groups may learn which member shows greater metacognitive awareness and subsequently defers the confidence judgments to that member (Duncan-Reid & McCarley, 2017). This strategy is akin to the *behavior and feedback* model posed by Bahrami et al. (2010), wherein group members defer Type 1 judgments to the most accurate member, which is equivalent to the groups treating the task as a disjunctive

one. In this behavior and feedback model, group performance is identical to that of the better member, regardless of the discrepancy in performance between the group members. Although the results of the present experiment found no credible difference between the groups and the better members in metacognitive efficiency, neither did it find any credible difference between the groups and the worse members. In observing the posterior distributions of metacognitive efficiency across the groups and the better and worse members, the group metacognitive efficiency distribution appears to fall between that of the worse members and the better members. It is therefore uncertain whether group metacognitive efficiency was equivalent to, or worse than, the better members.

In research by Bahrami et al. (2010), empirical data from collaborating pairs in a signal detection task was instead most consistent with a *weighted confidence sharing* model, wherein both the binary signal detection decision, and the confidence in that decision is communicated by the group members and weighted together in the final decision. In this weighted confidence sharing model, contrary to the behavior and feedback model, groups have the potential for collaborative gains but can show performance worse than their better members if there is a discrepancy between the sensitivities of the group members. By extending the weighted confidence sharing model from Type 1 sensitivity, to Type 2 sensitivity, this strategy could explain the observed suboptimality of group metacognitive efficiency. If groups combining and weighting their confidence judgments showed inefficiencies similar to those that plague Type 1 sensitivity judgments (assuming equal weights erroneously, avoiding unequal weights due to equality bias), the metacognitive efficiency of groups should

equate to the average of the worse and better members. To participants, both members making an independent confidence rating and simply making the group decision the average of the two may have seemed like an easy and perfectly reasonable strategy. If working collaboratively did not grant any increases in metacognitive awareness for the participants, this average confidence weighting strategy could explain the suboptimal metacognitive awareness of groups compared to ideal.

Group members might also use inefficient weighting strategies not because of defaulting to equal weights or a desire to avoid the embarrassment of acknowledging a member is comparatively worse than another, but because their assessment of their own performance changes when comparing themselves to others. The Dunning-Kruger effect refers to the phenomenon where individuals with poor abilities also show poor awareness of this lack of ability; individuals who are incompetent are not aware of their incompetence (Kruger & Dunning, 1999). This effect can occur in the opposite direction, and individuals who are highly skilled or show high performance can underestimate their own abilities relative to others. The poorer performance of the metacognitively worse group members may have prevented them from recognizing the ability and metacognitive awareness of the better member within in their groups as being superior to their own. When given the opportunity to compare their own performance to the performance of superior peers, individuals with the poorest ability do not revise their self-assessments even after this direct social comparison (Kruger & Dunning, 1999). The opposite is true for highly competent individuals, who instead better calibrate their perceived level of competence after comparison to others. Therefore, in the present study, the lack of a collaborative gain in group

metacognitive efficiency above that of the better members might be explained in-part by this effect. The metacognitively worse group members might fail to improve their metacognitive awareness during collaboration despite social comparison with a better member, nor notice any changes in their performance as a group relative to their own individual performance after the addition of a teammate superior to themselves. It might even be possible that metacognitively better individuals encounter difficulty in persuading a metacognitively worse individual to adopt a more ideal weighting strategy to improve both Type 1, and Type 2 performance for the group.

Limitations and Future Directions

One limitation of this study is that the 'groups' consisted of only 2 individuals, limiting the generalizability of the present findings. Pairs of individuals is the smallest possible group size, and in everyday life groups can consist of many more people. Although many studies of group task performance use only pairs (Bahrami et al., 2012b; Hertz et al., 2015; Mahmoodi et al., 2015), many also explore larger group sizes (Hinsz, 1990; Sniezek & Henry, 1989; Sorkin et al., 2001). Replicating this experiment with larger group sizes would explore whether collaborative gains, or further collaborative deficits, can be observed in this specific signal detection task with increasingly greater numbers of collaborators.

Prior to data collection and analysis, no method of assessing null hypotheses was planned or pre-registered in this study. While the present research can show that metacognitive efficiency did not differ credibly between the better members and the groups, it cannot claim that there is no difference without a Bayesian approach of suggesting evidence in favour of the null hypothesis. In future replications of this

research a region of practical equivalence, or Savage-Dickey methods of approximating the Bayes factor, should be utilized in the hierarchical models and planned a priori.

The short duration of this study (approximately 40 minutes completing trials) may have adversely affected the ability of groups to achieve collaborative gains. Compared to other studies of performance using similar gauge tasks in a signal detection framework (Montgomery, 1999; Sorkin et al., 2001), this study had relatively few trials and less overall time spent completing the task. The experiment of Sorkin et al. (2001) tested participants over multiple testing sessions and thousands of trials, as well as repeatedly assessed groups consisting of the same members, giving participants extensive experience working the same collaborators. The ability to perform the task efficiently, as well as discern the abilities of a group member and develop ideal weighting strategies may develop over time, or requires extensive experience with a teammate to emerge, and it is therefore a possibility that collaborative gains could not be observed in the short duration of the present experiment.

A final limitation of the present research is that while individuals made their own signal/noise judgments and confidence ratings individually, they only performed a single group decision when working collaboratively. This meant that it was impossible to examine how an individual's own performance and metacognition changed during collaboration. It might be possible that a better individual shows an increase in their metacognitive awareness when working with another person, but that increase is hidden after it has been weighted/integrated with the worse members

judgments in a single joint decision. Future research might then aim to include a method of private individual assessments alongside a final group decision weighted by both member's judgments, in order to isolate how the metacognition of individuals changes under collaborative conditions.

This study can endorse Maniscalco and Lau's (2012) *meta-d′* and metacognitive efficiency measures as useful tools to examine metacognition within signal detection tasks. Many different conceptualizations of metacognition exist, and with the *meta-d′* measure being an objective assessment of metacognitive ability, and the applicability and usefulness of signal detection theory, future research could aim to replicate past findings and research that use outdated and conceptually flawed correlational measures such as gamma and Pearson's r to assess the relationship between performance and metacognition.

Conclusion

When individuals collaborate in cognitive tasks, their combined group effort has the potential to exceed what even the best member of the group could achieve on their own. Past research has found many cases where groups show collaborative gains above that of single individuals, but most fall groups fall short of ideal predictions, and many perform worse than artificially groups individuals who do not interact. An area of research that has received renewed attention is metacognition, which involves the monitoring of one's own performance, cognitions, and strategies. It was of interest in the present experiment how metacognitive awareness might change under collaborative conditions.

The present study suggested that in a gauge-aggregation signal detection task, groups not only failed to show any credible collaborative gains in task performance, but they also failed to show any collaborative gains in metacognitive efficiency. In fact, group metacognitive efficiency showed a trend towards being worse than the metacognitively better members of the groups. Poor quality of verbal discussion, information integration, and weighting of group member contributions may have contributed to the observed suboptimality of collaborative metacognition.

Future research should aim to explore if group metacognitive efficiency changes vary after greater periods of collaboration, and with more experience with the specific gauges task. Similarly, larger groups, beyond pairs of 2 people should be examined to generalize these findings beyond such small groups. Finally, future research could examine the private decisions and judgments made by the individual members while collaborating, in addition to a single group judgment.

Bibliography:

Allen, M., Frank, D., Schwarzkopf, D. S., Fardo, F., Winston, J. S., Hauser, T. U., & Rees, G. (2016). Unexpected arousal modulates the influence of sensory noise on confidence. *Elife, 5*, e18103.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012a). Together, slowly but surely: the role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance, 38*(1), 3.

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012b). What failure in collective decision-making tells us about metacognition. *Philosophical transactions., 367*(1594), 1350-1365. doi:10.1098/rstb.2011.0420

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329*(5995), 1081-1085. doi:10.1126/science.1185718

Beran, M. J., Perner, J., & Proust, J. (2012). *Foundations of metacognition*: Oxford University Press.

Desender, K., Buc Calderon, C., Van Opstal, F., & Van den Bussche, E. (2017). Avoiding the conflict: Metacognitive awareness drives the selection of low-demand contexts. *Journal of Experimental Psychology: Human Perception and Performance, 43*(7), 1397.

Diehl, M., & Stroebe, W. (1991). Productivity loss in idea-generating groups: Tracking down the blocking effect. *Journal of personality and social psychology, 61*(3), 392.

Duncan-Reid, J., & McCarley, J. S. (2017, September). Collaborative Metacognition in a Signal Detection Task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1590-1593). Sage CA: Los Angeles, CA: SAGE Publications.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist, 34*(10), 906.

Garofalo, J., & Lester, F. K. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education, 16*(3), 163-176. doi:10.2307/748391

Hamilton, K., Mancuso, V., Mohammed, S., Tesler, R., & McNeese, M. (2017). Skilled and unaware: The interactive effects of team cognition, team metacognition, and task confidence on team performance. *Journal of Cognitive Engineering and Decision Making, 11*(4), 382-395.

Harkins, S. G. (1987). Social loafing and social facilitation. *Journal of experimental social psychology., 23*(1), 1-18. doi:10.1016/0022-1031(87)90022-9

Henry, R. A. (1993). Group judgment accuracy: Reliability and validity of postdiscussion confidence judgments. *Organizational behavior and human decision processes, 56*(1), 11-27.

Hertz, U., Romand-Monnier, M., Kyriakopoulou, K., & Bahrami, B. (2015). Social Influence Protects Collective Decision Making From Equality Bias. *Journal of Experimental Psychology*, *42*(2), 164.

Hinsz, V. B. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of personality and social psychology*, *59*(4), 705.

Ingham, A. G., Levinger, G., Graves, J., & Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of experimental social psychology*, *10*(4), 371-384.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology*, *65*(4), 681.

Kerr, N. L., & Tindale, R. S. (2004). Group Performance and Decision Making. *Annual Review of Psychology*, *55*(1), 623-655. doi:10.1146/annurev.psych.55.090902.142009

Kravitz, D. A., & Martin, B. (1986). Ringelmann rediscovered: The original article. *Journal of Personality and Social Psychology*, *50*(5), 936-941.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573-603. doi:10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*: Academic Press.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of personality and social psychology, 37*(6), 822.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., . . . Frith, C. D. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences, 112*(12), 3835-3840.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition, 21*(1), 422-430.

Maniscalco, B., McCurdy, L. Y., Odegaard, B., & Lau, H. (2017). Limited cognitive resources explain a trade-off between perceptual and metacognitive vigilance. *Journal of neuroscience, 37*(5), 1213-1224.

Montgomery, D. A. (1999). Human sensitivity to variability information in detection decisions. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 41*(1), 90-105.

Montgomery, D. A., & Sorkin, R. D. (1996). Observer sensitivity to element reliability in a multielement visual display. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 38*(3), 484-494.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological bulletin, 95*(1), 109.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review, 84*(3), 231.

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and cognition, 28*, 151-160.

Paulus, P. B., & Dzindolet, M. T. (1993). Social influence processes in group brainstorming. *Journal of personality and social psychology, 64*(4), 575.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods, 162*(1), 8-13.

Ringlemann, M. (1913). *Recherches sur les moteurs animés: Travail de l'homme*.

Shaw, M. E. (1980). *Group dynamics, the psychology of small group behavior*. New York: McGraw-Hill.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational behavior and human decision processes, 43*(1), 1-28.

Sorkin, R. D., & Dai, H. (1994). Signal Detection Analysis of the Ideal Group. *Organizational behavior and human decision processes, 60*(1), 1-13. doi:10.1006/obhd.1994.1072

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological review, 108*(1), 183-203.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers, 31*(1), 137-149.

Steiner, I. D. (1966). Models for inferring relationships between group size and potential group productivity. *Behavioral Science, 11*(4), 273-283.

Steiner, I. D. (1972). Group processes and group productivity. *New York: Academic*.

Watson, G. B. (1928). Do groups think more efficiently than individuals? *The Journal of Abnormal and Social Psychology, 23*(3), 328.

Weldon, M. S., & Bellinger, K. D. (1997). Collective memory: collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1160.

Wells, K., & Bradley, D. A. (2012). A review of X-ray explosives detection techniques for checked baggage. *Applied Radiation and Isotopes*, *70*(8), 1729-1746.

Appendix A.

Table A.1

Means [and Bayesian Credible Intervals] for d', meta-d', and metacognitive efficiency.

Condition	d′	Meta-d′	Metacognitive Efficiency
Worse Members	2.70 [2.54, 2.87]	1.68 [1.35, 2.01]	0.62 [0.50, 0.75]
Better Members	3.04 [2.88, 3.20]	2.48 [2.17, 2.81]	0.82 [0.71, 0.93]
Groups	3.23 [3.06, 3.39]	2.38 [2.07, 2.69]	0.74 [0.64, 0.84]