Received: 2 August 2017 | Revised: 9 September 2019 | Accepted: 18 September 2019

DOI: 10.1002/bdm.2156

RESEARCH ARTICLE

Collective decision making reduces metacognitive control and increases error rates, particularly for overconfident individuals

Matthew D. Blanchard | Simon A. Jackson | Sabina Kleitman

School of Psychology, The University of Sydney, Sydney, NSW, Australia

Correspondence
Matthew D. Blanchard, School of Psychology, The University of Sydney, Griffith Taylor Building (A19), Sydney, NSW 2006, Australia.
Email: matthew.blanchard@sydney.edu.au

Abstract

This research aimed to investigate the changes in judgment accuracy, confidence, control thresholds, and decision outcomes when people act in two-person groups (dyads) compared with acting individually. First, we used interacting dyads to determine the metacognitive and behavioral outcomes of collective decision making and compared them with those of individuals. Second, we examined whether these changes were related to the trait-confidence and bias of individuals working together. Using a within-person design, undergraduate psychology students ($N = 116$) completed a General-knowledge Test individually, then together as a dyad. Each question was accompanied by a confidence rating and a decision to bet $10 on the answer. Dyads had significantly higher confidence and lower control thresholds than individuals. They were also significantly more decisive (made more bets) and reckless (lost a higher rate of bets) than when working alone. Thus, we observed a higher rate of decision errors for groups than individuals. The results also demonstrated the important role of individual differences: Overconfident individuals became even more confident, decisive, and reckless when working together compared with less confident or underconfident individuals working together. These findings have important theoretical and applied implications for collective decision making; metacognitive bias and potentially control thresholds may be targeted to alleviate the larger error rates and guide the formation of more effective groups.

KEYWORDS

confidence, consensuality, decision-making tendencies, dyads, group decision making, metacognitive control, overconfidence

1 | INTRODUCTION

Group decision making is an everyday occurrence that we often trust because we believe “two heads are better than one” (Sunstein & Hastie, 2015). Acceptance of this axiom is observed through the widespread use of juries in legal systems, committees in governments and universities, and teamwork in organizations. Supporting this assumption, empirical studies have often demonstrated that group decisions can lead to more accurate judgments than individual decisions (Bahrami et al., 2010; Bahrami et al., 2012a, 2012b; Bahrami, Didino, Frith, Butterworth, & Rees, 2013; Bang et al., 2014; Bang et al., 2017; Henry, 1993; Hill, 1982; Hinsz, 1990; Koriat, 2012a, 2015; Laughlin, 2011; Laughlin, Bonner, & Miner, 2002; Mahmoodi et al., 2015; Massoni & Roux, 2017; Michaelsen, Watson, & Black, 1989; Sniezek & Henry, 1989; Tindale, 1989; Trouche, Sander, & Mercier, 2014; Wahn, Czeszumski, & Konig, 2018; Yaniv, 2011; Zarnoth & Sniezek, 1997). However, the question remains: Does this effect generalize to any group composition and to all types of problems?

Metacognitive confidence is often implicated as an important mechanism in the “two heads are better than one” effect as groups tend to select judgments proposed by the most confident member (Bahrami et al., 2010; Henry, 1993; Koriat, 2012a, 2015; Sniezek &

Henry, 1989). This strategy results in higher confidence for groups than individuals (Koriat, 2015; Patalano & LeClair, 2011; Savadori, Van Swol, & Sniezek, 2001; Sniezek & Henry, 1989; Zarnoth & Sniezek, 1997); however, it is not always associated with higher judgment accuracy (Heath & Gonzalez, 1995; Minson & Mueller, 2012; Schuldt, Chabris, Woolley, & Hackman, 2017). Additionally, the size of a group's increase in confidence depends on the trait-confidence of each group member: Lower trait-confidence members had the largest increase in confidence when working collectively compared with alone (Schuldt, Chabris, Woolley, & Hackman, 2017). This finding has broader implications, especially in regard to the calibration of confidence in the group. In this research, we propose that the accuracy of one's confidence judgments (indexed by bias) will also be associated with the size of a group's increase in confidence.

Moreover, there is a scarcity of research that has assessed metacognitive control processes in collectively made decisions. Control processes regulate the amount of cognitive effort and time invested in the accumulation of evidence that informs decisions (Ackerman, 2014; Jackson, Kleitman, Howie, & Stankov, 2016; Jackson, Kleitman, Stankov, & Howie, 2016; Jackson, Kleitman, Stankov, & Howie, 2017; Koriat & Goldsmith, 1996). Research on individuals has shown that control thresholds relate to decision-making errors (Jackson & Kleitman, 2014; Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017; Koriat & Goldsmith, 1996; Pansky, Goldsmith, Koriat, & Pearlman-Avnion, 2009). We will examine this relationship for groups and whether collective control thresholds differ from individual thresholds.

There is also a paucity of research examining the patterns of decision behaviors that groups engage in, as research typically assesses the accuracy of judgments only. Jackson and colleagues developed an individual differences framework that allows the computation of behavioral measures from the outcomes of decisions (Jackson et al., 2017; Jackson & Kleitman, 2014; Jackson, Kleitman, Stankov, & Howie, 2016). This work revealed that confidence was associated with decision error rates: Individuals with higher confidence were more decisive (placed more bets on their judgments being correct) and more reckless (lost a higher rate of bets) than individuals with lower confidence. This framework has yet to be applied to group decision making.

The aim of this research was to integrate and extend previous work by investigating the behavioral outcomes of group decisions that are associated with metacognitive control and confidence. Specifically, we look at whether groups, which tend to be more confident than individuals, differ in their control thresholds, make more erroneous decisions (higher recklessness), and whether the magnitude of error rates depend on individual differences in trait-confidence and/or bias.

1.1 | Confidence theory

Bahrami's et al. (Bahrami et al., 2010) seminal work on the "two heads better than one" effect provided an important extension of earlier work that demonstrated the influential role of confidence on decision making (e.g., Hinsz, 1990; Littlepage, Schmidt, Whisler, & Frost, 1995; Sniezek & Van Swol, 2001; Yaniv, 1997; Zarnoth & Sniezek, 1997). Participants were briefly shown two visual stimuli, each containing six identical patterns. The luminance was increased for a single pattern within one of the stimuli to produce a nonconforming target. Participants judged which stimuli contained the nonconforming target by themselves and then together with an allocated partner. When working as a dyad, group members could interact freely and had to agree on a joint response. The joint responses were found to be more accurate than the responses given by participants working alone. The best explanation for the observed data was that participants shared and used each other's subjective confidence as a proxy for the probability that they were correct. We will refer to this as the confidence theory. This is a plausible explanation because confidence and accuracy tend to be positively correlated (Koriat, 2008, 2012b; Yaniv, 1997). Crawford and Stankov (1998) demonstrated that the magnitude of the relationship varied with the type of task: The average correlation was .43 for measures of fluid intelligence and .35 for measures of crystallized intelligence (see Stankov, 1999).

This finding received robust support from Koriat (2012a, 2015). Koriat (2015) had participants judge which of two lines was longer in one study and answer two alternative general-knowledge questions in a second study. Both tests were completed twice: first on their own and then together with a partner. They also gave separate confidence ratings (ranging from 50% to 100%) for their individual responses and joint responses. Importantly, Koriat included "Consensually Wrong" (CW) questions (sometimes referred to as "misleading" or "deceptive," see Brewer & Sampaio, 2012; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994) for which the wrong answer is chosen by the majority of people and, usually, with higher confidence than the correct answer. CW questions contrast with "Consensually Correct" (CC) questions for which most people select the correct answer with higher confidence than the wrong answer. Koriat's two studies replicated the "two heads *better* than one" effect when examining CC questions. However, a "two heads *worse* than one" effect was found for CW questions.

Koriat (2015) also applied a maximum confidence slating algorithm to test the theory that confidence drives these effects. This algorithm creates virtual dyads by selecting the individual decision of the most confident group member for each question, thus modeling what should occur if groups use confidence to guide their decisions. The same pattern of results was observed for the virtual dyads and real dyads. Real dyads generally adopted the individual response associated with higher confidence, leading to improved group accuracy on CC items but poorer accuracy on CW items, and increased group confidence in both cases.

By untangling confidence and accuracy, Koriat (2015) provided robust empirical evidence for the confidence theory that decisions tend to be made via shared feelings of confidence between group members. He demonstrated that group responses were given more confidently than individual responses regardless of their accuracy and in alignment with the responses made by the most confident group member. These findings reveal the conditions under which accuracy will benefit or be harmed by "two heads."

Group confidence also appears to be influenced by consensus: When group members initially agree on a response, their confidence tends to be higher than when they initially disagree (Budescu & Rantilla, 2000; Budescu, Rantilla, Yu, & Karelitz, 2003; Budescu & Yu, 2007; Koriat, 2015; Sniezek & Buckley, 1995; Yaniv, Choshen-Hillel, & Milyavsky, 2009). It is unclear whether group confidence aligns with the most confident member under both consensus conditions (i.e., agreement and disagreement) because the few existing studies provide mixed results. Sniezek and Buckley found that group confidence was higher than individual confidence for agreement but not disagreement trials, whereas group confidence was higher for both agreement and disagreement trials in Koriat's study.

The present study employed CC and CW questions to disentangle the influence of accuracy and confidence. We compared individual and group decision-making outcomes for both types of questions.[1] We also used the maximum confidence slating algorithm to create virtual dyads that model the outcomes that should occur when groups rely on confidence to guide their decision making. Then we compared the outcomes of virtual dyads and real dyads to determine whether any changes in group decision-making outcomes can be attributed to either confidence or accuracy.

1.2 | Confidence and decision making

Confidence is a critical mechanism that guides decisions under uncertainty; judgments held with higher confidence tend to have a greater behavioral impact (Aramovich & Larson, 2013; Fischhoff, Slovic, & Lichtenstein, 1977; Koriat, 2011; Koriat & Goldsmith, 1996; Koriat, Lichtenstein, & Fischhoff, 1980; Pansky & Goldsmith, 2014). For example, Koriat and Goldsmith (1996) found that the decision to submit an answer for marking on a General-knowledge Test was associated with significantly higher confidence than the decision to withhold it. The result was the same regardless of whether the decision was a hit (i.e., correct answer) or false alarm (i.e., wrong answer).

This work has been extended to an individual differences paradigm (Jackson & Kleitman, 2014; Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017), in which the rates of various decisions (hits, correct rejections, false alarms, and misses) were converted into patterns of decision outcomes. Although this method resembles a signal detection approach, there is an important distinction: Our decision outcome variables only take frequencies into account, whereas signal detection variables (e.g., d') incorporate additional parameters related to variance. These parameters impose distributional assumptions on environmental cues that do not apply to monitoring cues underlying decision behavior (Fetsch, Kiani, & Shadlen, 2014; Koriat & Goldsmith, 1996; Maniscalco & Lau, 2012; Ratcliff & Starns, 2013). The two behavioral decision outcomes relevant to the present study were decisiveness and recklessness. Decisiveness indicates the tendency to initiate action regardless of the accuracy of the associated judgment. For example, if offered the opportunity to bet on the correctness of an answer, a highly decisive individual should place a greater proportion of bets (i.e., hits and false alarms) than a less decisive individual. Recklessness refers to the tendency to initiate poor decisions or how often their decisiveness is misguided (i.e., false alarms only). For example, a highly reckless individual should lose more bets than an individual with lower recklessness. Jackson and colleagues have repeatedly found that individuals with higher confidence are more decisive and reckless than those with lower confidence (Jackson & Kleitman, 2014; Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017).

This research investigated an important applied implication of behavioral decision-making research and the confidence theory of group decisions: selecting the more confident member's response would lead to increased dyadic confidence and, as a result, increased error rates (i.e., recklessness) for dyads compared with participants working alone. That is, dyads would be more decisive (i.e., place more bets on their responses being correct) and more reckless (i.e., lose a higher proportion of bets when their judgments are wrong) decision makers than individuals.

1.3 | Control and decision making

Decision errors also relate to metacognitive control processes that regulate the amount of cognitive effort and time invested in making decisions (Ackerman, 2014; Jackson, Kleitman, Howie, & Stankov, 2016, Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017; Koriat & Goldsmith, 1996). When making decisions, individuals tend to use confidence to monitor the process of accumulating evidence for a choice (Alter & Oppenheimer, 2009; Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1999; Koriat, 2012b; Pleskac & Busemeyer, 2010). The control threshold is the minimum level of confidence required to make a decision or take some action (Jackson, Kleitman, Howie, & Stankov, 2016; Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017; Koriat & Goldsmith, 1996). For example, a medical clinician may decide to administer treatment to a patient only when they are 90% confident that their diagnosis is correct (e.g.,Djulbegovic et al., 2014 ; Pauker & Kassirer, 1980). That is, they would administer treatment only when they believe they would be correct nine times out of 10. Jackson and colleagues computed the threshold level of confidence required for individuals to submit their answers for marking on a cognitive test or administer treatment to fictive patients on a medical decision-making test (Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017). They referred to the control threshold as the Point Of Sufficient cerTainty (POST). They found that individuals with lower POSTs had higher rates of decision errors (i.e., higher recklessness). They also found that the level of one's POST was not related to the average level of their confidence judgments. That is, an individual who tends to respond with high confidence could have a high, low, or moderate POST. To our knowledge, no research has investigated POSTs at the group level, so it is unclear whether group POSTs differ from individual POSTs. If confidence increases when working in a

[1] We reran the analyses to include consensus: agreement and disagreement trials. The results are presented in Appendix B.

group compared with alone, groups should increase their POSTs to a similar degree to avoid increasing errors rates (recklessness). However, we predicted that working in a group would be associated with higher error rates (recklessness) than working alone, so we expected that group POSTs would change (i.e., difference between individual and group responses) at a different rate than confidence. The magnitude of this change is unknown as no research has yet examined the difference between individual and group POSTs.

1.4 | The role of individual differences in trait-confidence

The existence of individual differences in levels of confidence is also relevant to this inquiry. Regardless of their nature, when different cognitive tests are utilized with confidence ratings (e.g., general-knowledge, reasoning, and visual), a broad confidence factor defined by confidence judgments on each test emerges in addition to accuracy factors defined by accuracy scores. That is, relative to others, some people generally have low confidence judgments, whereas others generally have high confidence judgments. This tendency is robust and has been replicated in numerous studies (e.g., Kleitman & Gibson, 2011; Kleitman & Stankov, 2001, 2007; Pallier et al., 2002; Soll, 1996; Stankov, Lee, Luo, & Hogan, 2012; Stankov, Pallier, Danthiir, & Morony, 2012). Thus, this factor is commonly referred to as trait-confidence (see Stankov, Kleitman, & Jackson, 2014 for review). The finding that confidence is trait-like has implications for its measurement in decision-making research. This domain has often measured confidence using an alternative version of the same experimental test (e.g., Schuldt et al., 2017), but this ignores the robust nature of trait-confidence. A measure of individual trait-confidence obtained using multiple cognitive tests will tend to be more reliable than a single measure. Additionally, as these tests are different to the experimental test used to measure dyadic outcomes, the likelihood of issues such as collinearity and statistical dependency are expected to decrease. We measured trait-confidence using two cognitive tests that were not used in the experimental manipulation.

A question addressed in the current research is how differences in trait-confidence influence group decision making. Rather than investigating the contribution of these individual differences, prior experimental studies have often removed it as a source of variance by standardizing confidence judgments (e.g., Koriat, 2012a). Although such elimination might be useful in testing theoretical hypotheses, individual differences are likely to affect the quality of real-life group decisions. A recent study suggested that the trait-confidence levels of group members influence collective decision making. Schuldt et al. (2017) found that when two individuals with low trait-confidence were paired together, the level of confidence in their joint responses increased considerably. For two individuals with mixed trait-confidence (i.e., one high and one low), dyadic confidence increased moderately. For two high trait-confidence individuals, there was no difference in confidence between individuals and dyads. The confidence theory suggests that this finding has implications for the behavioral outcomes (i.e., decision tendencies: decisiveness and recklessness) of collectively made decisions.

In this study, we investigated whether the magnitude of a dyad's increase in decisiveness (overall proportion of bets) and recklessness (false alarm rate) depends on individual differences in trait-confidence. On the basis of Schuldt et al.'s (2017) findings, we expected that dyads composed of members with lower trait-confidence would demonstrate the largest increases in decisiveness and recklessness. Furthermore, we measured individual levels of trait-confidence using two cognitive tests that were different from the experimental test completed by dyads.

1.5 | The role of individual differences in bias

We also expected the accuracy of one's confidence judgments (e.g., calibration) to be associated with decisiveness and recklessness. Calibration relates to the adaptiveness and effectiveness of the self-monitoring process (Nelson, 1996; Stankov, 1999) and reveals the appropriateness of confidence ratings given an individual's level of accuracy (Keren, 1991; Yates, 1990). It is trait-like (albeit, a construct separate from trait-confidence) and also varies greatly between individuals (Kleitman, 2008; Schraw, Dunkle, Bendixen, & Roedel, 1995; Stankov & Crawford, 1996, 1997). As such, calibration behaves differently to trait-confidence: Individuals with the same level of trait-confidence but different levels of accuracy will have divergent levels of calibration. Individual differences in calibration have potential implications for the formation of accurate group judgments; however, there remains limited research in the group decision-making literature (e.g., Pescetelli, Rees, & Bahrami, 2016). Massoni and Roux (2017) found that groups composed of members with diverse levels of calibration (i.e., an overconfident and an underconfident member) made more decision errors than group members with similar levels of calibration (i.e., both members overconfident or underconfident). The authors suggested that diverse groups made more errors because the lower performing member's decisions were endorsed by the group more often than was justified by their ability. Given that judgments held with higher confidence tend to have a greater behavioral impact (Aramovich & Larson, 2013; Fischhoff et al., 1977; Koriat et al., 1980; Koriat & Goldsmith, 1996) and that overconfident individuals have inflated beliefs about their competence on a task (Stankov & Crawford, 1996, 1997), we expected that the effect of calibration similarity on recklessness would depend on whether members were underconfident or overconfident.

This research investigated whether the size of a dyad's increase in decisiveness and recklessness depended on individual differences in the accuracy of confidence ratings, measured using the bias score. Similar to calibration, bias captures one's ability to accurately self-monitor their performance on a task and indicates the degree of overconfidence or underconfidence in one's judgments on average: Positive scores indicate overconfidence, negative scores indicate underconfidence, and scores approaching zero (±10 percentage units) indicate unbiased confidence ratings (Keren, 1991; Stankov et al., 2014; Yates, 1990). We expected decisiveness and recklessness to increase the most for overconfident (e.g., both members overconfident) compared with underconfident (e.g., both members underconfident) or unbiased (e.g., both members unbiased, or one underconfident member and one overconfident member) dyads. These findings carry important practical implications for the selection of group members and the reduction of decision errors.

1.6 | Control variables

Additional variables were assessed to control for their potentially confounding influence. They include cognitive abilities (e.g., Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017; Jackson, Kleitman, Howie, & Stankov, 2016; Rudolph, Niepela, Greiffa, Goldhammerb, & Kröner, 2017), Big-Five personality factors (Intellect, Conscientiousness, Extraversion, Agreeableness, and Neuroticism; e.g., Jackson & Kleitman, 2014; Jackson, Kleitman, Stankov, & Howie, 2016; Lauriola & Irwin, 2001), gender (e.g., Estes & Hosseini, 1988; Powell & Ansic, 1997), age (e.g., Worthy, Gorlick, Pacheco, Schnyer, & Maddox, 2011), group member familiarity (e.g., Harrison, Mohammed, McGrath, Florey, & Vanderstoep, 2003), and English language ability. Controlling for these variables ensures that our results can be linked to trait-confidence and bias.

1.7 | Aims and hypotheses

The overarching goal was to investigate the nature of changes in judgment accuracy, confidence, control thresholds (estimated using POSTs), and decision outcomes when people acted in dyads compared with when they acted individually. The primary aim was to examine whether group decisions were more erroneous than individual decisions and if so, whether the confidence theory accounts for these increases. The secondary aim was to determine whether group-level decision-making changes could be explained by the trait-confidence or bias of group members in real dyads. To investigate these aims, we used three approaches: (a) Virtual dyads, derived from empirical data, were calculated utilizing a maximum confidence slating algorithm (Bang et al., 2014; Koriat, 2012a, 2015; Pescetelli et al., 2016); (b) the POSTs and behavioral decision-making outcomes of real dyads were compared with those of individuals and virtual dyads; and (c) determine whether the magnitude of changes in collective decision-making outcomes depended on individual differences in trait-confidence or bias after controlling for cognitive ability, personality, gender, age, group member relationship, and English language ability.

We expected to replicate previous findings that real dyads correctly answer more CC questions but fewer CW questions than individuals. Similarly, the answers of real dyads were expected to be associated with higher confidence than individual answers. Our novel hypotheses are as follows:

Hypothesis 1a. *POSTs will change at a different rate than confidence, when responding as a dyad compared with alone.*

Hypothesis 1b. *Dyads will bet more frequently on the correctness of their answers than individuals, regardless of questions being CC or CW (i.e., dyads will be more decisive).*

Hypothesis 1c. *Dyadic bets will have a greater false alarm rate than individual bets, regardless of questions being CC or CW (i.e., dyads will be more reckless).*

Hypothesis 2a. *The increase in betting behavior (Hypothesis 1a) and false alarm rate (Hypothesis 1b) for dyads will depend on the trait-confidence of its members. Following Schuldt et al. (2017), the magnitude of these changes will be greatest for dyads composed of lower trait-confidence members.*

Hypothesis 2b. *The increase in betting behavior (H1a) and false alarm rate (H1b) for dyads will be higher for dyads composed of more overconfident members.*

2 | METHOD

2.1 | Participants

In return for partial course credit, 116 Australian undergraduate psychology students participated in the study (88 female, M_{age} = 20.11, *SD* = 4.34, range 16–48). Six were excluded for not having a dyad partner, and two were excluded for not completing the experiment. Two participants did not follow instructions and were identified as outliers with unrealistic responses. One of these participants bet on ~90% of their general-knowledge responses being correct, and the other assigned excessively high individual confidence ratings (i.e., more than three standard deviations above the mean) on the test. Thus, these participants and their partners were excluded prior to all analyses.[2] The final sample involved 52 dyads (*n* = 104, 80 female, M_{age} = 20.08, *SD* = 4.57, range 16–48). Of these, 66% were born in Australia, 72% spoke English as a first language, and 6% knew their dyad partner prior to participation.

2.2 | Experimental tasks

2.2.1 | General-knowledge Test

This test composed 30 general-knowledge questions that participants judged as "True" or "False" (Brewer & Sampaio, 2012; Schuldt et al., 2017). Of these, 25 were selected from a larger list used in previous research investigating confidence changes in dyads (Schuldt et al.,

[2] We conducted all analyses reported in Section 3 with the two identified outliers included in the dataset. The results were the same as what is reported.

2017). For example, "The bicycle was invented in Scotland (T*/F)." The five remaining questions were taken from research reporting them as CW questions (Brewer & Sampaio, 2012). We aimed for an equal split of CC and CW questions to reduce measurement error and achieve acceptable reliability estimates for the metrics derived from the General-knowledge Test (i.e., accuracy, confidence, decisiveness, and recklessness) at both levels of question consensuality (see Cortina, 1993; Cronbach, 1951; Komorita & Graham, 1965; Lord & Novick, 1968). Half of the 30 questions were True, and the other half were False. The questions were presented in a fixed order, and participants answered each question twice: first individually and then together with their partner. Schuldt et al. (2017) reported excellent test–retest reliability for confidence (.84) but poor reliability for accuracy (−.19). No other study has investigated POSTs, decisiveness, or recklessness using these general-knowledge questions. Our reliability estimates are described in Section 3.

2.2.2 | Esoteric Analogies Testl

This test required participants to complete 18 analogies (Stankov, 1997). On the basis of an original pair of words, participants chose one of four alternatives that shared the same relationship with a target word. For example, "LOVE is to HATE as FRIEND is to: (1) LOVER (2) PAL (3) OBEY (4) ENEMY*?" Accuracy requires both reasoning ability and acquired knowledge. The 18 questions were selected from the original 24 on the basis of previous research with Australian undergraduate students (Jackson, Kleitman, Howie, & Stankov, 2016), which has reported acceptable internal consistency estimates for accuracy (.74) and confidence (.94). Our reliability estimates were also acceptable (see Table 3).

2.2.3 | Raven's Advanced Progressive Matrices

This test included 12 items that each displayed a 3 × 3 matrix of abstract figures following a horizontal and vertical pattern (Raven, 1938–1965). The bottom right figure was blank, requiring participants to choose which of eight alternatives completed the pattern. Accuracy in this test is a measure of abstract reasoning ability. The 12 items were selected from the original 36 on the basis of previous research with Australian undergraduate students (Jackson, Kleitman, Howie, & Stankov, 2016), which has reported acceptable internal consistency estimates for accuracy (.68) and confidence (.84). Likewise, our reliability estimates were acceptable (see Table 3).

2.2.4 | Mini-IPIP

This questionnaire asked participants to rate the degree to which 20 statements accurately described them using a 5-point rating scale (Donnellan, Oswald, Baird, & Lucas, 2006). For example, participants rated "Am the life of the party" from being a *very inaccurate* (1) to *very accurate* (5) description of them. This scale measures the Big-Five personality factors and possesses acceptable internal reliability for Agreeableness (.70), Conscientiousness (.69), Extraversion (.77), Intellect (.65), and Neuroticism (.68; Donnellan et al., 2006). Our reliability estimates were acceptable (see Table 3).

2.3 | Procedure

Participants attended the lab in groups of two to six and were randomly allocated to dyads upon arrival. Dyad partners were seated next to each other. Computers were separated by partitions, so members could see each other but not the other's screen. Room dividers separated different dyads, so they could not see or easily hear each other. After providing consent, participants could opt in to win a double movie voucher for the best performing individual and the best performing dyad based on accuracy on the General-knowledge Test. Participants then completed all measures in the same order: demographics, General-knowledge Test, Esoteric Analogies Test, Mini-IPIP, and Raven's Advanced Progressive Matrices. All measures were completed individually except for the General-knowledge Test. Figure 1 shows the sequence for each question on the General-knowledge Test. Participants answered a question alone and then again with their dyad partner before moving onto the next question. When responding individually, a general-knowledge question appeared on each member's screen. Participants indicated a response using their keyboard. Participants then entered how confident they were that their answer was correct, between 50% (guessing) and 100% (completely confident). Then they were asked if they would place a hypothetical $10 bet on their answer being correct. Responses were indicated using their keyboard. Participants were instructed to wait for their partner before proceeding to the group phase. When both dyad members were ready, they pressed the spacebar on their keyboards. The same general-knowledge question appeared on each member's screen accompanied by instructions to "Discuss your answer with your partner. Try to persuade them if necessary. Come to an agreement and give the same answer." After submitting the same answer, participants indicated how confident they were that their group answer was correct. They were instructed to answer this alone, without discussing their level of confidence with their partner. They were then asked if they would place a hypothetical $10 bet on their answer being correct with instructions to "Discuss your decision with your partner. Try to persuade them if necessary. Come to an agreement and make the same decision." Participants entered the same decision into their computers, and the next question began.

2.4 | Calculation of variables

Using each participant's responses when working alone and in a group, we calculated mean accuracy, confidence, POST (calculation described in detail in Appendix A), decisiveness (proportion of bets placed), and recklessness (proportion of bets lost) across all general-

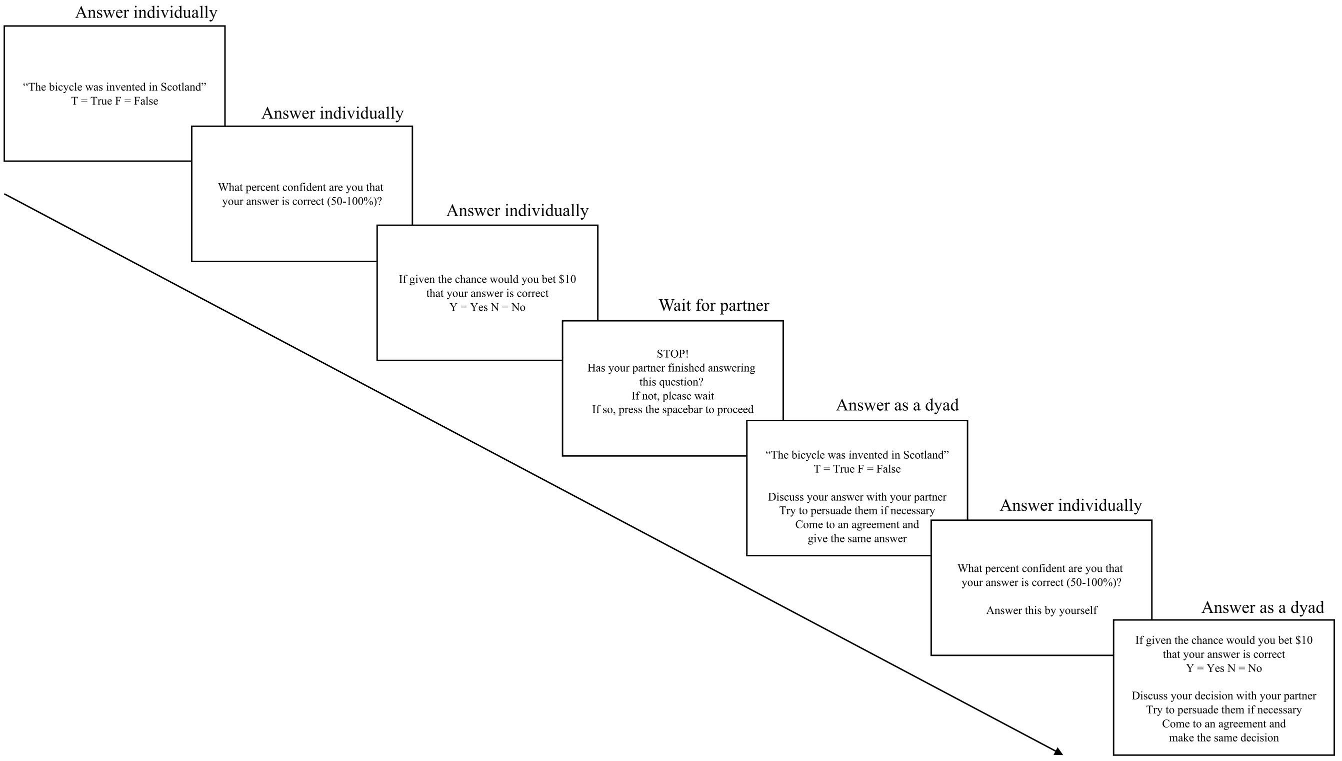
Answer individually
"The bicycle was invented in Scotland"
T = True F = False
Answer individually
What percent confident are you that your answer is correct (50-100%)?
Answer individually
If given the chance would you bet $10 that your answer is correct
Y = Yes N = No
Wait for partner
STOP!
Has your partner finished answering this question?
If not, please wait
If so, press the spacebar to proceed
Answer as a dyad
"The bicycle was invented in Scotland"
T = True F = False
Discuss your answer with your partner
Try to persuade them if necessary
Come to an agreement and give the same answer
Answer individually
What percent confident are you that your answer is correct (50-100%)?
Answer this by yourself
Answer as a dyad
If given the chance would you bet $10 that your answer is correct
Y = Yes N = No
Discuss your decision with your partner
Try to persuade them if necessary
Come to an agreement and make the same decision

FIGURE 1 Order of question presentation for the General-knowledge Test

knowledge questions for individuals, real dyads, and virtual dyads. A summary of the calculations for these variables is described in Table 1.

For the regression analyses, we calculated each dyad's change in accuracy, confidence, POST, decisiveness, and recklessness from working alone to working together. These dyad–individual difference scores were calculated as the mean dyad score (e.g., mean dyadic accuracy) minus the mean individual score (e.g., mean individual accuracy). Consistent with previous research (see Stankov et al., 2014 for a review), we calculated trait-confidence and bias. It should be noted that trait-confidence was calculated from the mean of confidence ratings on Esoteric Analogies Test and Raven's Advanced Progressive Matrices, and bias from the mean of confidence ratings minus the percentage of questions answered correctly on Esoteric Analogies Test. We used Esoteric Analogies Test for bias because it is a mixed marker of crystallized and fluid intelligence, which is more appropriate for performance on a General-knowledge Test than Raven's Advanced Progressive Matrices.

3 | RESULTS

3.1 | Descriptive statistics

Consistent with prior research (Bahrami et al., 2010; Bahrami et al., 2012b; Bang et al., 2014; Koriat, 2012a, 2015; Schuldt et al., 2017), analyses were based on dyads as the unit of analysis, meaning "individual" results refer to the average of the dyad members working alone. Table 2 reports the descriptive statistics and reliability estimates for the General-knowledge Test variables. Mean accuracy and confidence were comparable with prior research using the same test (Schuldt et al., 2017). As they did not investigate POSTs or patterns of decision outcomes (decisiveness and recklessness), there is no prior research to compare these variables. Internal consistency estimates were acceptable for all variables except accuracy, the POST, and individual recklessness. The lack of internal consistency for accuracy indicated that the General-knowledge Test did not capture a stable trait related to accuracy such as crystallized intelligence. This is *consistent*

TABLE 1 Summary of calculations for variables derived from the General-knowledge Test

Variable	Calculation
Accuracy	Percentage of questions answered correctly
Confidence	Mean of all confidence ratings
POST	Level of confidence required to place a bet on the correctness of an answer. Calculated using logistic regression (see Appendix A for detailed description)
Decisiveness	Number of decisions to place bets/total number of decisions
Recklessness	Number of decisions to place bets following incorrect judgments (false alarms)/total number of decisions following incorrect judgments

Abbreviation: POST, Point Of Sufficient cerTainty.

with Schuldt et al. (2017) who reported poor reliability for individual accuracy using questions from the same pool. Low internal consistency occurred for individual POST for CC questions (.51), dyadic POST for CW questions (.43), and individual recklessness for CC (.45) and CW questions (.35). Internal consistency was just acceptable for dyadic recklessness for CC (.57) and CW questions (.61). These lower reliabilities may have occurred because the calculation of POSTs and recklessness was based on a smaller subset of the data. POSTs were calculated for either CC or CW questions, and recklessness was calculated for decisions to place bets following incorrect judgments only. Additionally, internal consistency for recklessness may have been lower because it was derived, in part, from accuracy.

General-knowledge question consensuality was determined post hoc. As defined by Koriat (2008, 2012b), a question was classified as CC if more than 50% of individuals answered it correctly or CW if less than 50% of individuals answered it correctly. As planned, there was almost an even split, with 13 questions classified as CC and 17 as CW.[3]

Following a recent trend in the group decision-making literature, we calculated within-person gamma correlations to further investigate the relationship between accuracy and confidence for the two types of questions (i.e., CC and CW; Bang et al., 2014; Koriat, 2012a, Koriat, 2015; Nelson, 1984). The gamma correlations are described in detail in Appendix B and suggest that participants were not aware of the deceptive nature of CW questions.

We checked the frequency of bets placed for individuals and real dyads to ensure that participants took the bet decisions seriously and did not bet on every answer regardless of their probability of being correct. Figure 2 shows the frequency of bets across the 30 general-knowledge questions. The number of bets placed ranged from 0 to 22 for individuals and 1 to 22 for real dyads. Given this distribution and that no individual or dyad placed bets on all questions, we believe participants took the bet decisions seriously and tended to place bets based on their subjective assessment of being correct.

Table 3 reports the descriptive statistics and reliability estimates of the control variables. They are comparable with similar studies using Australian undergraduate samples (e.g., Jackson, Kleitman, Stankov, & Howie, 2016; Jackson & Kleitman, 2014; Kleitman, 2008; Kleitman & Stankov, 2007).

3.2 | Analysis of variance

Hypotheses 1a, 1b, and 1c were examined via a series of 2 × 2 repeated-measures analyses of variance (ANOVAs).[4] Figure 3 shows the mean accuracy, confidence, POST, decisiveness (overall proportion of bets), and recklessness (false alarm rate) on CC and CW questions for the General-knowledge Test when answered by individuals, real dyads, and virtual dyads.

3.2.1 | Individuals versus real dyads

We examined the POSTs and decision outcomes of real dyads compared with when the same participants acted individually. The independent variables were grouping (individuals vs. real dyads) and question consensuality (CC vs. CW).[5] Table 4 reveals the results of this series of ANOVAs.

For accuracy, there was no difference between individuals and real dyads, but as expected, accuracy was greater for CC than CW questions. The interaction was not significant; thus, we failed to replicate the "two heads *better* than one" or "two heads *worse* than one" effects.

For confidence, real dyads were more confident than individuals, and confidence was greater for CC than CW questions. The interaction term was significant; thus, we conducted simple effects tests (reported below) to interpret this effect. Confidence was significantly higher for real dyads than individuals for both CC, mean difference = −3.43, $t(51) = -8.74$, $p < .001$, $d = -0.68$, and CW questions, mean difference = −2.68, $t(51) = 6.56$, $p < .001$, $d = -0.62$, indicating that real dyads had higher confidence than individuals, and this effect was more pronounced for CC than CW questions.

For the POST, real dyads had significantly lower POSTs than individuals, regardless of question consensuality. POSTs did not differ for CC or CW questions, and the interaction was not significant. To examine whether the change in POSTs corresponded with the change in confidence, we compared the difference between individual responses and group responses on confidence with the difference on the POST. For CC questions, the difference on the POST (mean = 1.40) was significantly lower than the difference on confidence, mean = −3.43, $t(49) = -5.90$, $p < .001$, $d = 1.08$. For CW questions, the difference on the POST (mean = 1.93) was significantly lower than the difference on confidence, mean = −2.68, $t(48) = -4.62$, $p < .001$, $d = 0.88$. These results indicate that dyadic POSTs did not change (from individual POSTs) at a similar rate to confidence.

For decisiveness, real dyads were significantly more decisive than individuals, and decisiveness was significantly greater for CC than CW questions. The interaction was not significant.

For recklessness, real dyads were significantly more reckless than individuals, and there was no difference between CC and CW questions. The interaction was also nonsignificant.

[3] Surprisingly, two of the five CW questions we used from Brewer and Sampaio (2012) were CC in our sample; we categorized them according to our results. We reanalyzed the data with these two questions removed, and the findings were mostly the same. The only difference was for Hypothesis 2a; trait-confidence significantly predicted the change in recklessness. That is, higher trait-confidence individuals became even more reckless compared with lower trait-confidence individuals when working together in a dyad.

[4] We were unable to combine these analyses into one set of four 3 × 2 ANOVAs because of the dependency between responses of individuals and virtual dyads.

[5] We randomly removed 10 CW questions (creating a 65/35 split of CC/CW questions) to check that our results were not influenced by the number of CW questions. The results were the same as those reported in this section.

TABLE 2 Descriptive statistics and reliability estimates for the General-knowledge Test variables

	Individual			Dyad		
Variable	Mean	*SD*	RE	Mean	*SD*	RE
General-knowledge Test						
Accuracy (% correct)						
Total	48.77	6.60	.03	49.38	7.92	−.23
CC	62.91	9.02	−.06	63.59	12.23	−.18
CW	37.97	8.78	−.03	38.49	12.37	.07
Confidence (mean %)						
Total	61.21	3.95	.88	64.21	4.74	.90
CC	62.79	4.72	.78	66.22	5.40	.81
CW	60.01	3.77	.78	62.68	4.80	.84
POST (mean %)						
Total	69.24	6.73	.78	67.57	6.39	.82
CC	68.41	7.48	.51	67.23	7.08	.73
CW	69.44	6.96	.58	67.56	6.88	.43
Decisiveness (% of bets placed)						
Total	26.86	12.92	.81	38.51	18.33	.82
CC	31.76	15.23	.68	45.56	21.38	.66
CW	23.13	12.69	.67	33.16	19.21	.74
Recklessness (% of bets lost)						
Total	27.28	14.40	.60	35.51	19.64	.64
CC	25.95	21.53	.45	34.92	29.89	.57
CW	27.58	13.12	.35	35.50	19.84	.61

Note. Means and standard deviations (*SDs*) were calculated using dyads as the unit of analysis. RE is the internal reliability estimate, which was computed using Cronbach's α for all variables except the POST and recklessness. For the POST and recklessness, internal reliability was estimated by correlating the odd and even questions corrected by the Spearman–Brown prophecy formula (Guilford, 1954; Jackson et al., 2017; Stankov & Crawford, 1996). Reliability estimate coefficients were calculated using individuals as the unit of analysis.
Abbreviations: CC, Consensually Correct; CW, Consensually Wrong; POST, Point Of Sufficient cerTainty.

3.2.2 | Virtual dyads versus real dyads

We created virtual dyads to compare what theoretically should occur if groups rely on confidence to what was observed for the real dyads. The results of these analyses are described in detail in Appendix C. In summary, the outcomes for real dyads were consistent with those of the virtual dyads, indicating that real dyads tended to make decisions that were consistent with the most confidence dyad member.

3.3 | Consensus (agreement vs. disagreement)

Consensus tends to be related to group confidence (Budescu et al., 2003; Budescu & Rantilla, 2000; Budescu & Yu, 2007; Koriat, 2015; Sniezek & Buckley, 1995; Yaniv et al., 2009), suggesting that it may also be associated with the behavioral outcomes of group decisions under investigation (i.e., decisiveness and recklessness). We reran the series of ANOVAs with consensus (agreement vs. disagreement) as an additional variable in each model. An agreement trial occurred when both dyad members entered the same individual response for a general-knowledge question, and a disagreement trial occurred when they entered different individual responses. The results were mostly the same and will only be summarized here (see Appendix D for a detailed description). On agreement trials, dyads had significantly higher confidence, decisiveness, and recklessness and lower POSTs than individuals, regardless of question consensuality. On disagreement trials, dyads had significantly higher confidence and decisiveness and lower POSTs than individuals, regardless of question consensuality. For recklessness, on disagreement trials, there was no significant difference between individuals and dyads for CC or CW questions. There was also no significant difference between individuals and dyads on accuracy for agreement or disagreement trials. The results revealed that the effects of working in a dyad were more pronounced when dyad members agreed on a true–false response than when they disagreed.

3.4 | Regression analyses

Hypotheses 2a and 2b were tested via a series of simultaneous regression analyses. The dependent variables were dyad–individual

TABLE 3 Descriptive statistics and reliability estimates for the control variables

Variable	Mean	*SD*	RE
Accuracy (% correct)			
EAT	70.08	10.68	.59
RAPM	47.76	18.50	.80
General cognitive ability	58.92	12.49	.78
Confidence (mean %)			
EAT	71.79	9.18	.88
RAPM	60.86	15.25	.90
Trait-confidence	66.32	10.39	.90
Bias	1.71	11.42	.61
Personality (mean)			
Agreeableness	3.93	0.46	.67
Conscientiousness	3.07	0.57	.69
Extraversion	3.13	0.63	.85
Intellect	3.75	0.54	.67
Neuroticism	2.88	0.50	.62

Note. RE is the internal reliability estimate, which was computed using Cronbach's α for all variables. Reliability estimate coefficients were calculated using individuals as the unit of analysis. Means and standard deviations (*SDs*) were calculated using the dyad-level dataset. Abbreviations: EAT, Esoteric Analogies Test; RAPM, Raven's Advanced Progressive Matrices.

difference scores for accuracy, confidence, POST, decisiveness (overall proportion of bets), and recklessness (proportion of bets lost). The dyad–individual difference score represented each dyad's change from members working alone to working together: A positive score indicated an increase for dyads, a negative score indicated a decrease for dyads, and zero represented no change. The predictor variables of interest were trait-confidence and bias. As noted earlier, the predictor variables were calculated from Esoteric Analogies Test and Raven's Advanced Progressive Matrices, not the General-knowledge Test. Correlations between trait-confidence and bias and the dyad–individual difference variables are presented in Table 5.

Except for accuracy and the POST, all the dyad–individual difference scores correlated significantly and positively with trait-confidence (*r* = .29 to.31, *p* < .05) and bias (*r* = .37 to.48, *p* < .05). We also considered several control variables capturing demographic characteristics (gender, age, group member familiarity, and English language ability), cognitive ability, and Big-Five personality factors. Prior to regression analyses, we examined the correlations between all variables (see Table E1). Of the control variables, only Extraversion and Neuroticism shared significant relationships with any of the dyad–individual difference scores. Thus, they were also included in the regression analyses.[6] Detailed correlations are provided in Appendix E.

Table 6 presents the regression results of each dyad–individual difference variable being regressed on trait-confidence, Extraversion, and Neuroticism (Hypothesis 2a).

In the overall regression, trait-confidence was the only significant predictor of the dyad–individual difference scores for confidence, *F*(3, 48) = 4.41, *p* = .04, and decisiveness, *F*(3, 48) = 4.92, *p* = .03. Trait-confidence uniquely explained 7.8% to 8.3% of the variance in these variables, respectively. Trait-confidence was not a significant predictor of the dyad–individual difference for accuracy, *F*(3, 48) = 0.41, *p* = .52, the POST, *F*(3, 45) = 2.69, *p* = .11, or recklessness, *F*(3, 48) = 4.03, *p* = .05. It is worth noting that recklessness was just beyond the cutoff for significance, and the amount of unique variance (7%) trait-confidence accounted for in recklessness was comparable with confidence and decisiveness. Further, neither Extraversion nor Neuroticism was a significant predictor of any of the dyad–individual difference scores (controlling for trait-confidence). Figure 4 shows the scatter plots and resulting regression slopes between the dyad–individual difference variables and trait-confidence.

Table 7 presents the regression results of each dyad–individual difference variable being regressed on bias, Extraversion, and Neuroticism (Hypothesis 2b).

In the overall regression, bias was the only significant predictor of the dyad–individual difference scores for confidence, *F*(3, 48) = 10.49, *p* < .01, decisiveness, *F*(3, 48) = 4.64, *p* = .04, and recklessness, *F*(3, 48) = 7.38, *p* < .01. Bias uniquely explained 18.21%, 8.63%, and 13.33% of the variance in these variables, respectively. Bias was not a significant predictor of the dyad–individual difference for accuracy, *F* (3, 48) = 0.11, *p* = .74, or the POST, *F*(3, 45) = 1.84, *p* = .18. In addition, Extraversion and Neuroticism did not significantly predict any of the dyad–individual difference scores (controlling for bias). Figure 5 shows the scatter plots and regression slopes between the dyad–individual difference variables and bias.

The findings suggest that trait-confidence and bias of dyad members influence collective decision making. When working together, the greatest increases in confidence and decisiveness (overall proportion of bets) occurred for dyads composed of higher trait-confidence members, but recklessness did not reach significance. The effect of trait-confidence was, however, in the opposite direction to our prediction and the findings of Schuldt et al. (2017). Thus, Hypothesis 2a was not supported. We also found that dyads composed of more overconfident members demonstrated the greatest increase in confidence, decisiveness (overall proportion of bets), and recklessness (proportion of bets lost). Thus, Hypothesis 2b was supported. Trait-confidence and bias had no effect on accuracy changes.

4 | DISCUSSION

The present study investigated the changes in the POST and patterns of behavioral decision-making outcomes that occurred when individuals made decisions alone compared with as a dyad. Our results support the confidence theory and provide further evidence that group decision making is guided by subjective confidence. Due to their

[6] We also conducted the regression analyses for Hypotheses 2 without the control variables (Extraversion and Neuroticism) in the models. The findings were mostly the same; however, one difference occurred. Trait-confidence significantly predicted the magnitude of the dyadic change in recklessness. Higher trait-confidence individuals became even more reckless compared with lower trait-confidence individuals when working together in a dyad.

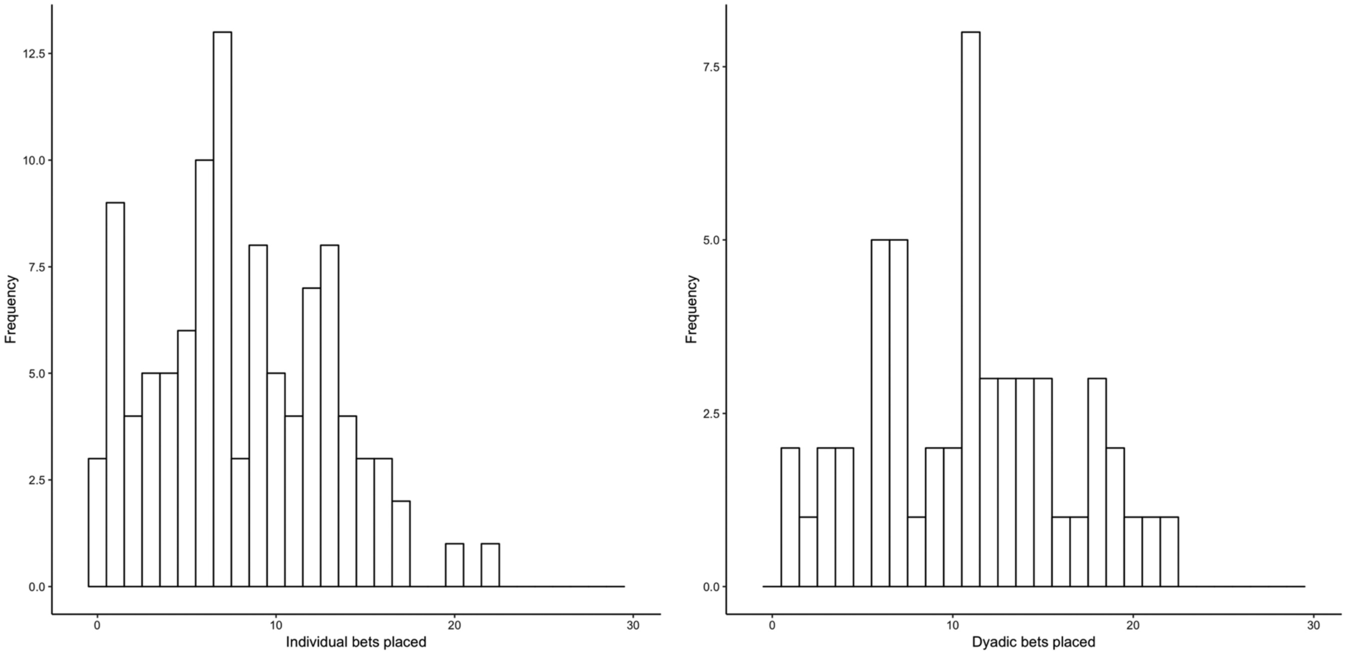

FIGURE 2 The frequency of bets placed by individuals and dyads on the General-knowledge Test (out of 30 possible bets)

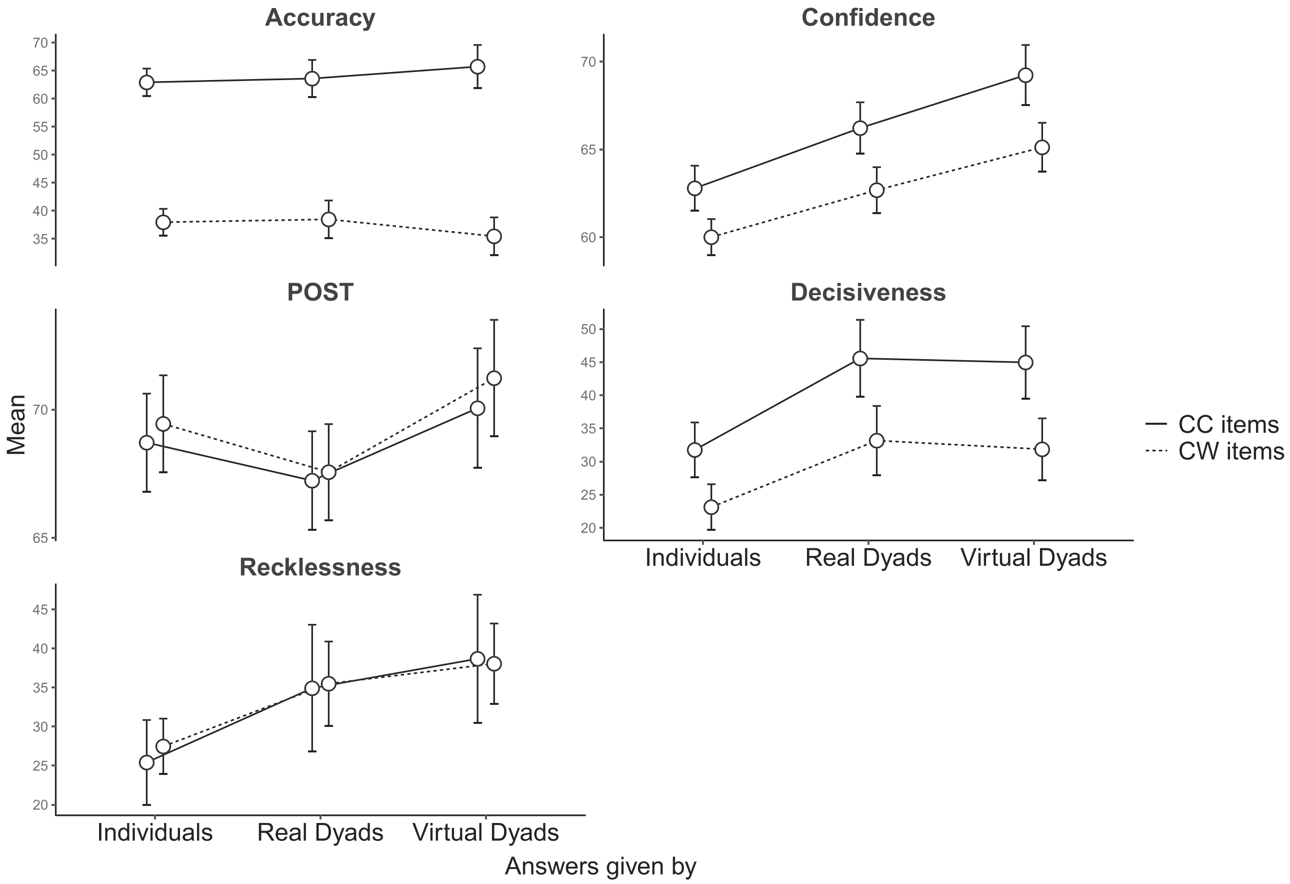

FIGURE 3 Mean score for general-knowledge test variables depending on grouping (Individuals vs. Real Dyads vs. Virtual Dyads) and question consensuality (CC vs. CW). Errorbars represent 95% confidence intervals

TABLE 4 Individuals versus dyads: Series of 2 × 2 analyses of variance of the difference between grouping (individual vs. real dyad) and question consensuality (Consensually Correct vs. Consensually Wrong) on accuracy, confidence, decisiveness, and recklessness

	Mean difference	*F* (1, 51)	*p*	η^2	η_p^2
Accuracy					
Grouping	−0.60	0.38	.54	–	–
Consensuality	25.02	157.85	<.001***	.58	.76
Grouping × Consensuality	–	0.01	.93	–	–
Confidence					
Grouping	−3.00	72.88	<.001***	.10	.59
Consensuality	3.16	56.98	<.001***	.10	.53
Grouping × Consensuality	–	4.40	.04*	.002	.08
POST					
Grouping	1.68	6.21	.02*	.01	.07
Consensuality	0.85	1.91	.17	–	–
Grouping × Consensuality	–	0.00	.98	–	–
Decisiveness					
Grouping	−11.65	63.94	<.001***	.11	.56
Consensuality	10.51	45.11	<.001***	.08	.47
Grouping × Consensuality	–	2.97	.09†	–	–
Recklessness					
Grouping	−8.23	15.62	<.001***	.04	.23
Consensuality	−1.06	0.29	.60	–	–
Grouping × Consensuality	–	0.16	.69	–	–

Note. Two dyads demonstrated no recklessness, so the degrees of freedom for this analysis were lower than the others (*df* = 1, 49). POSTs could not be estimated for all individuals or dyads, so the degrees of freedom were lower for this analysis: grouping (*df* = 1, 50) and consensuality (*df* = 1, 47).
Abbreviation: POST, Point Of Sufficient cerTainty.
***p < .001.*p < .05.†p < .10.

TABLE 5 Correlations between the dyad–individual difference variables and trait-confidence and bias

Dyad–individual difference variable	Trait-confidence	Bias
Accuracy	−.08	−.03
Confidence	.29*	.48***
POST	.21	.23
Decisiveness	.31*	.37**
Recklessness	.29*	.39**

Abbreviation: POST, Point Of Sufficient cerTainty.
***p < .001.**p < .01.*p < .05.

tendency to rely on confidence and lower POSTs, collectively made decisions had higher error rates than decisions made by individuals. Combining different methodologies, we found that the magnitude of these group-level changes tended to increase as the trait-confidence (calculated using different tests) and overconfidence (positive bias score) of group members increased.

4.1 | Accuracy and confidence

Prior research has demonstrated the conditions under which “two heads” benefit and harm the performance of interacting dyads (Koriat, 2012a, 2015). As no accuracy differences were observed between individuals and real dyads for CC or CW questions, we did not replicate these findings. This was unexpected given that these effects were observed for virtual dyads. It appears that our failure to replicate the two heads are *better* and *worse* than one effects occurred because groups did not always select the response of the most confident member. Schuldt et al. (2017), who used the same pool of general-knowledge questions, also found no accuracy difference between individuals and real dyads, but they did not investigate question consensuality. It is worth noting that unlike much of prior research, we employed a difficult General-knowledge Test that may offer insight into why these effects were not replicated. Structural features of the decision-making context may also have contributed to this null result. For example, the opportunity for groups to improve accuracy above the individual level tends to occur for CC questions when members initially disagree. Given that 13 of 30 questions in our study were categorized as CC and group members initially disagreed on 45% of trials, these opportunities for improvement occurred for a small subset of trials. Additionally, when disagreement occurs, there is likely a greater level of uncertainty than agreement trials, potentially leading to lower confidence ratings. Thus, the confidence cue used by members to guide their decisions may be weaker. Taken together, these structural features

TABLE 6 Multiple regression analyses of the changes in dyadic responding as a function of trait-confidence and the control variables

	Dyad–individual difference variable									
	Accuracy		Confidence		POST		Decisiveness		Recklessness	
Predictor	*β*	sr^2	*β*	sr^2	*β*	sr^2	*β*	sr^2	*β*	sr^2
Trait-confidence	−.09	.01	.28*	.08	.23	.05	.29*	.08	$.27^\dagger$	.07
Extraversion	−.03	.00	$.25^\dagger$	.06	.14	.02	.20	.04	.04	.00
Neuroticism	−.16	.02	−.09	.01	.25	.06	−.21	.04	$-.27^\dagger$	.08
Total R^2		.03		.16*		.11		.20*		.17*
n	52		52		49		52		52	

Note. β represents a standardized regression coefficient; sr^2 represents the unique variance each independent variable accounts for in the dyad–individual difference variable after controlling for the effects of the other independent variables. This was calculated by squaring the semipartial (part) correlation (Cohen, Cohen, West, & Aiken, 2003).

Abbreviation: POST, Point Of Sufficient cerTainty.

$*p < .05$. $^\dagger p < .10$.

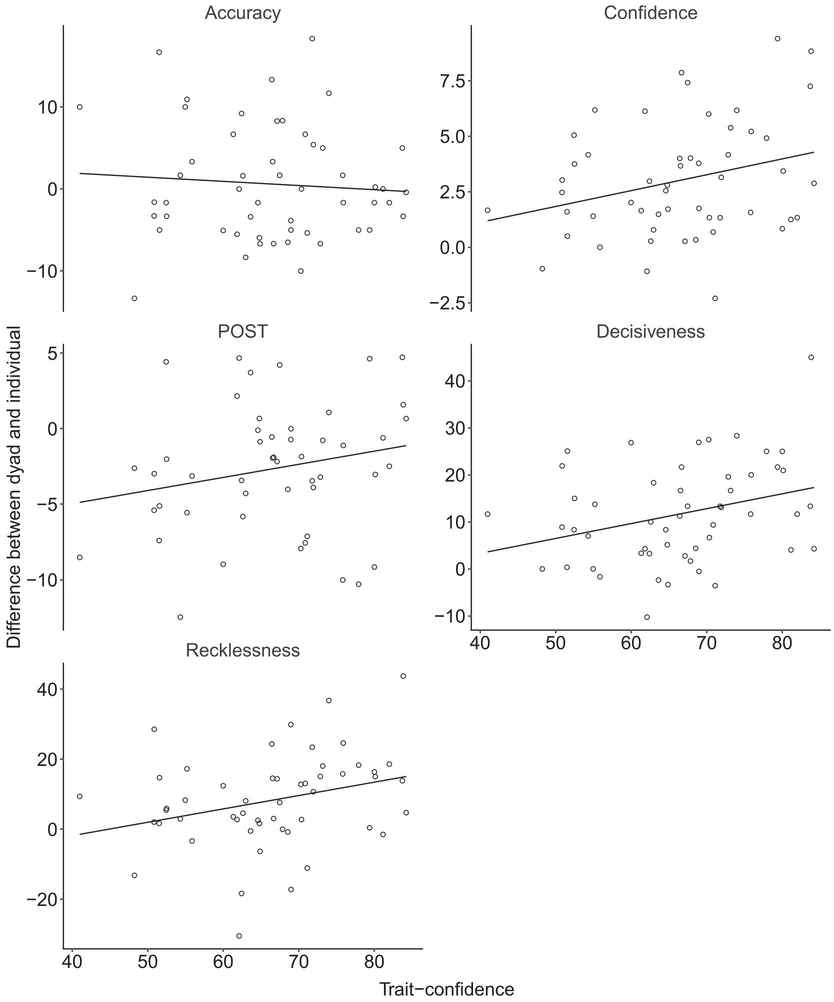

FIGURE 4 Change in mean accuracy, confidence, decisiveness, and recklessness depending on the trait-confidence of the dyad members

of the decision-making context may indicate a small effect size for accuracy that may require more questions and a larger sample to accurately detect.

The confidence theory proposes that people generally share and use each other's subjective confidence as a proxy for the likelihood that they are correct. As a result, the confidence levels of group members tend to align with the most confident member, making the group more confident overall. We observed this effect for both agreement and disagreement trials, replicating the finding that confidence is amplified for real dyads compared with the same participants working alone (Heath & Gonzalez, 1995; Koriat, 2015; Minson & Mueller, 2012; Patalano & LeClair, 2011; Schuldt et al., 2017; Sniezek & Henry, 1989; Zarnoth & Sniezek, 1997). However, the level of confidence of the real dyads did not increase to that

TABLE 7 Multiple regression analyses of the changes in dyadic responding as a function of bias and the control variables

	Dyad–individual difference variable									
	Accuracy		Confidence		POST		Decisiveness		Recklessness	
Predictor	β	sr^2	β	sr^2	β	sr^2	β	sr^2	β	sr^2
Bias	−.05	.00	$.43^{**}$	.18	.20	.04	.29*	.08	$.37^{**}$	.13
Extraversion	−.01	.00	.12	.02	.14	.02	.12	.01	−.06	.00
Neuroticism	−.15	.02	−.07	.01	.19	.04	−.21	.04	$-.26^{\dagger}$	.07
Total R^2		.02		$.25^{**}$		.10		.20*		$.22^{**}$
n	52		52		49		52		52	

Note. β represents a standardized regression coefficient; sr^2 represents the unique variance each independent variable accounts for in the dyad–individual difference variable after controlling for the effects of the other independent variables. This was calculated by squaring the semipartial (part) correlation (Cohen et al., 2003).

Abbreviation: POST, Point Of Sufficient cerTainty.

$^{**}p < .01$. $^{*}p < .05$. $^{\dagger}p < .10$.

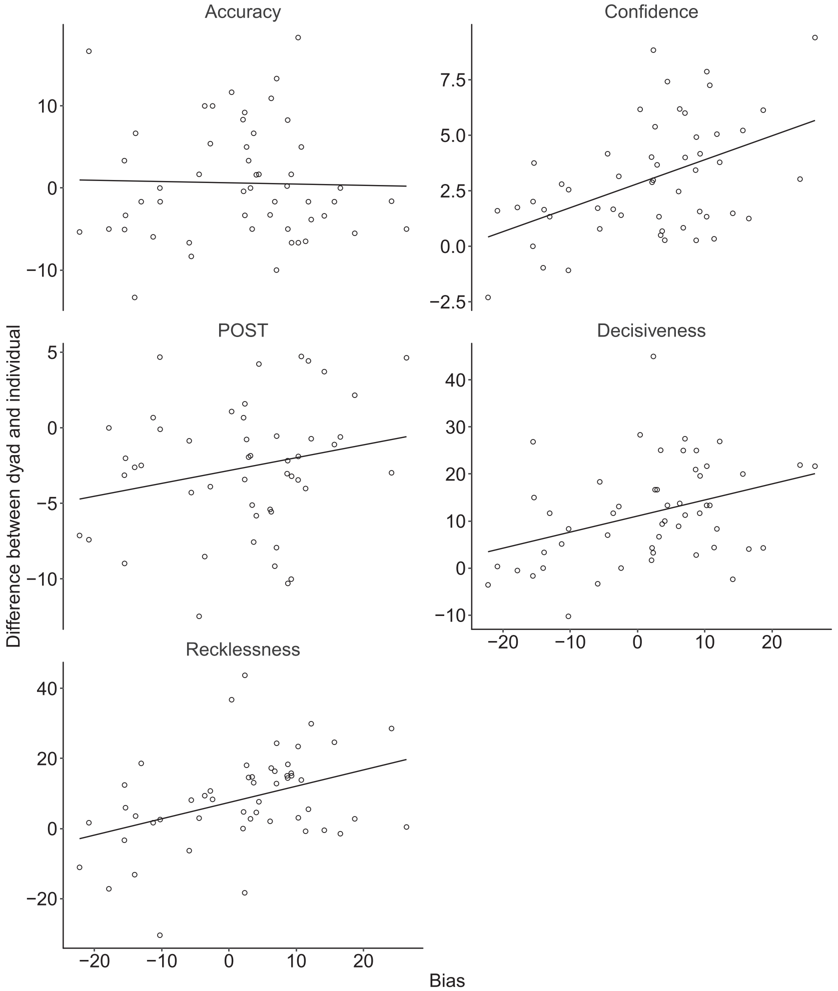

FIGURE 5 Change in mean accuracy, confidence, decisiveness, and recklessness depending on the bias of the dyad members

of the most confident group member. This likely occurred because groups did not always select the response of the more confident member for disagreement trials. This is consistent with Koriat's (2015) findings. Two possible explanations are that (a) high-quality argument may overcome the influence of confidence on some occasions (Trouche et al., 2014) and (b) discussion allows group members to work out differences in their levels of confidence. This second point may lead a group to endorse the less confident member's response because it is believed to have a higher probability of being correct (Bang et al., 2014).

4.2 | Control thresholds and confidence

We also observed that dyads tended to use lower control thresholds than individuals to guide their bet decisions and this change did not correspond with a dyad's change in confidence. These findings reveal that, compared with individuals, dyads tended to believe their general-knowledge answers had a higher probability of being correct (indicated by their greater confidence) and were willing to place bets when their levels of certainty about being correct were lower (indicated by their lower POSTs). This may suggest that dyads required less evidence than individuals to place bets on their answers being correct. Control thresholds are known to be sensitive to the payoff context of the decision environment (e.g., Ackerman, 2014; Aminoff et al., 2012; Jackson et al., 2017; Koriat & Goldsmith, 1996), so these results may have occurred because of the low-stakes context. Thus, we would not expect dyads to consistently set lower POSTs than individuals.

4.3 | Decision-making outcomes

Prior research has demonstrated that confidence is positively related and the POST is negatively related to decisiveness (overall proportion of bets) and recklessness (false alarm rate or bets lost) within and between individuals (Jackson & Kleitman, 2014; Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017; Koriat & Goldsmith, 1996; Pansky et al., 2009). Our study is the first to extend these findings to collective decision making using behavioral measures of decision-making outcomes. We found that real dyads were more decisive (i.e., made more bets) and reckless (i.e., lost more bets) than individuals. This effect occurred for both CC and CW questions, indicating that our findings were not influenced by the number of CW items and apply equally to nonmisleading contexts where uncertainty may be lower. We can attribute these findings to the influence of confidence rather than accuracy because the findings from the real dyads were consistent with the virtual dyads. When we included consensus (agreement vs. disagreement), with some minor caveats, the results were the same. The influence of working in a dyad was most pronounced when group members initially agreed on an answer compared with when they disagreed (see Appendix D for detailed results). Overall, the results suggest that collective decision making leads to higher error rates, which may be explained by the sharing, use, and therefore increased subjective confidence and potentially decreased control thresholds.

4.4 | The influence of trait-confidence and bias

We observed a robust effect that dyads composed of higher trait-confidence individuals, on average, increased in confidence and decisiveness (i.e., made more bets) more than lower trait-confidence individuals when working collectively compared with alone. In addition, we observed even stronger effects for dyads composed of more overconfident (positive bias score) individuals. These dyads had the largest increases in confidence, decisiveness, and recklessness (i.e., lost more bets) when working collectively compared with alone. Given that neither trait-confidence nor bias shared a meaningful relationship with accuracy, it appears that these relationships were largely driven by confidence. Several explanations are possible. For example, two high trait-confidence individuals might receive validation through discussion, further amplifying their confidence and betting behavior. Despite no increase in accuracy, these groups become (more) overconfident, required less certainty to place bets, and consequently lost a higher rate of bets. Conversely, two low trait-confidence members might remain uncertain despite discussion, therefore maintaining underconfidence or lower overconfidence and betting behavior. These findings partially conflict with prior research that found that groups composed of members with similar levels of bias perform better than group members with diverse levels of bias (Bang et al., 2014; Massoni & Roux, 2017). We found that the effect of similarity on group decision errors depends on whether members are underconfident or overconfident. Recklessness increased the least when both members were underconfident, most when both were overconfident, and somewhere in-between when members were both well calibrated or had different levels of bias. Overall, these results suggest that the output of collective decision making is dependent on the types of individuals working together.

Nevertheless, we are hesitant to make strong claims about the trait-confidence result because it is the reverse direction of what was predicted based on Schuldt et al.'s (2017) findings. It is unclear why our results differed in direction; one reason may be the different methods for measuring the self-confidence trait, which is typically measured using cognitive tests (Kleitman & Gibson, 2011; Stankov et al., 2014; Stankov & Crawford, 1997). Schuldt et al. used the mean of individual confidence ratings from an alternative version of the same General-knowledge Test to represent trait-confidence. As shown in both Schuldt et al. and the current study, the General-knowledge Test was not capturing cognitive ability. We measured trait-confidence as the mean of confidence ratings from two cognitive measures that were not used in the experimental manipulation (i.e., Esoteric Analogies Test and Raven's Advanced Progressive Matrices) and have been used extensively in trait-confidence research (Jackson & Kleitman, 2014; Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017; Kleitman, 2008; Stankov et al., 2014). Thus, our measure of trait-confidence more closely aligns with theory and empirical work. Another possible explanation for the divergent results relates to the levels of confidence expressed by participants in the two studies. In the study conducted by Schuldt et al. (2017) compared with the present study, mean confidence was markedly higher for individuals (69.70 vs. 61.21) and especially for dyads (74.10 vs. 64.21). This difference suggests a possible ceiling effect in Schuldt et al.'s study such that high confidence group members already expressed their confidence at the upper end of the confidence range when responding individually; thus, they had little opportunity to express higher confidence when working together, particularly when they initially agreed on the group judgment, whereas in our study, confidence was, on average, expressed at lower levels, and consequently high confidence members were able to increase their confidence further when working together. Regardless, the results need to be replicated.

It is worth noting that no control variables were found to be predictors of the degree to which dyads increased in confidence, decisiveness, or recklessness. Thus, they were omitted from the relevant discussions. However, Neuroticism correlated significantly and negatively with decisiveness and recklessness. This is consistent with a study examining indecisiveness using a self-report measure, but they also found meaningful relationships with Conscientiousness and Extraversion (Patalano & LeClair, 2011). In our sample, the relationship Neuroticism shared with decisiveness and recklessness lost significance once its effect was considered incrementally to trait-confidence and bias. This is congruent with previous research adopting the same behavioral decision-making measures (Jackson & Kleitman, 2014; Jackson, Kleitman, Stankov, & Howie, 2016), which has found no consistent relationships between the Big-Five personality dimensions and decisiveness or recklessness. Thus, the prediction that trait-confidence and bias had on dyad–individual difference scores cannot be attributed to cognitive abilities, personality, gender, or age (albeit a number of caveats are described in Section 4.6).

4.5 | Implications

Our findings indicate that overconfidence (positive bias score) may be responsible for the failure to improve group decisions. We found that, on average, dyads became more confident than individuals despite no increase in accuracy, required lower levels of certainty to place bets (i.e., lower POSTs), and as a result made more reckless decisions than individuals. With the exception of the control threshold, these changes were more pronounced for dyads composed of members with higher trait-confidence (except recklessness) and greater overconfidence. Thus, the utility of forming groups and trusting them to make superior choices should be evaluated depending on the context. Our findings indicate that metacognitive constructs (i.e., trait-confidence and bias) may be targeted to inform the selection of group members. For example, the presence of one cautious individual (i.e., lower trait-confidence and underconfident) may alleviate some of the unjustified increases in confidence, decisiveness, and recklessness displayed by dyads composed of overconfident and potentially higher trait-confidence individuals. This may be most important in contexts characterized by high uncertainty and inadequate knowledge. In such situations, decisive decision making may lead to harmful errors. Traditional conceptions of error detection tend to focus on monitoring (i.e., confidence), but the flexibility of the control threshold makes it a potential candidate for interventions aimed at reducing dyadic errors. Future research should investigate dyadic control thresholds in a range of different tasks and payoff contexts. Given that outcomes were the same for both nonmisleading (CC) and misleading (CW) questions, our findings are not dependent on the type of question and are generalizable to nonmisleading contexts where uncertainty is lower, and the correct answer may be clear. Thus, our results are relevant for guiding the effective formation of groups across a broad range of decision-making contexts. The results also support our argument that future decision-making research should employ a similar methodology to measure trait-confidence. This method is consistent with theory and empirical research (e.g., Kleitman & Gibson, 2011; Kleitman & Stankov, 2001, 2007; Pallier et al., 2002; Soll, 1996; Stankov, Lee, et al., 2012; Stankov, Pallier, et al., 2012) and would have the added benefit of reducing collinearity and statistical dependency issues. Lastly, our results demonstrate that control thresholds and patterns of behavioral decision outcomes (e.g., decisiveness and recklessness) are important for future research to consider alongside judgment accuracy and confidence.

4.6 | Limitations and future directions

Despite these promising implications, several limitations must be considered. First, we did not replicate the "two heads *better* than one" and "two heads *worse* than one" effects. One possibility is that the General-knowledge Test used in this study did not capture stable individual differences (as indicated by low internal consistency estimates), which may be linked to the inclusion of CW questions and the high level of difficulty observed for our pool of questions. It is important to consider this as a caveat when interpreting the results of this study and how it relates to prior work. Another possibility is that structural features of the decision-making context reduced our ability to detect a potentially small effect size. Future studies should employ a larger number of items and recruit a larger sample to address this.

Second, individuals were not deliberately paired based on their level of trait-confidence. Schuldt et al. (2017) established each individual's trait-confidence before inviting only high and low trait-confidence individuals to participate in the dyadic component of the study. Due to logistical constraints, it was not possible to pre-screen and deliberately pair participants; however, this is recommended for future research.

Third, consistent with previous research (e.g., Bahrami et al., 2012a; Bahrami et al., 2013; Bahrami et al., 2010; Bang et al., 2014; Henry, 1993; Koriat, 2015; Mahmoodi et al., 2015; Massoni & Roux, 2017; Pescetelli et al., 2016; Schuldt et al., 2017; Sniezek & Henry, 1989; Wahn et al., 2018; Zarnoth & Sniezek, 1997), we used a within-subjects design, meaning dyads completed each general-knowledge question twice: first individually and then together as a dyad. In the real world, dyads may only consider a problem once, together, before agreeing on a joint decision; thus, this type of design may confound the results. Future research should either compare the results for interacting and noninteracting dyads completing each question as individuals and dyads or adopt a between-subjects design to remove the possible influence of being presented with questions twice and functioning in both individual and group roles.

Fourth, it is unclear whether confidence is causal of the observed effects or a by-product of discussion. The inclusion of a non-interacting condition (as mentioned in the previous paragraph) would allow future studies to examine the effect of discussion on group confidence and control thresholds.

Fifth, due to constraints on recruitment, our sample only consisted of two-person groups that may limit the generalizability of our findings to larger groups. Prior research suggests that larger groups also rely on confidence when making decisions, so we would expect our findings to extend. For example, confidence also appears to be greater for triads than individuals (Sniezek & Henry, 1989). Interestingly, Zarnoth and Sniezek (1997) found that collective overconfidence decreased as group size increased, so our results may not generalize to all group sizes. Future research should investigate the decision outcomes of groups of varying sizes.

Further limitations relate to the sample used in this study. The university sample is largely homogeneous compared with the general population. The age and intelligence ranges are restricted (e.g., young, high cognitive ability). Prior research indicates that older people place less emphasis on confidence when making decisions (Pansky et al., 2009) and are less competent decision makers (see Del Missier, Mäntylä, & Nilsson, 2015 for review) compared with younger people. Furthermore, people with higher cognitive abilities tend to have better decision-making skills (Del Missier, Mäntylä, & Bruin, 2012) and make more optimal decisions (Jackson & Kleitman, 2014; Jackson, Kleitman, Stankov, & Howie, 2016; Jackson et al., 2017). Future research should seek to recruit a larger, more heterogeneous sample to investigate how these demographics influence decision-making processes and outcomes.

5 | CONCLUSION

The present research makes numerous novel contributions to the group decision-making literature. We integrated different approaches and theories that have previously worked in isolation to form a unified framework for investigating the behavioral decision-making outcomes of dyads and potentially larger groups. In doing so, we were the first to examine the control thresholds of dyads and patterns of dyadic decision-making tendencies. We observed that dyads placed a higher number of bets (i.e., more decisive) and lost a greater proportion of bets (i.e., more reckless) than individuals. These effects were associated with the higher levels of confidence and lower control thresholds experienced by dyads than individuals. These findings occurred for nonmisleading and misleading questions and agreement and disagreement trials, which suggests that our findings are generalizable to these contexts. We also demonstrated a robust method for calculating trait-confidence and showed that the size of a dyad's increase in confidence, decisiveness, and recklessness depended on individual differences in trait-confidence (with the exception of recklessness) and metacognitive bias. Thus, the rate of decision-making errors made by groups may be reduced by targeting control thresholds and deliberately pairing individuals based on their level of bias (and potentially trait-confidence), leading to improved group effectiveness.

ACKNOWLEDGEMENTS

The authors thank Asher Koriat for his valuable comments on early versions of this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

ORCID

Matthew D. Blanchard https://orcid.org/0000-0001-5557-8617
Sabina Kleitman https://orcid.org/0000-0001-5772-5019

REFERENCES

Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General*, *143*(3), 1349-1368. https://doi.org/10.1037/a0035098

Alter, A., & Oppenheimer, D. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Aminoff, E., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., … Miller, M., (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition*, *40*(7), 1016–1030. https://doi:10.3758/s13421-012-0204-6

Aramovich, N. P., & Larson, J. R. (2013). Strategic demonstration of problem solutions by groups: The effects of member preferences, confidence, and learning goals. *Organizational Behavior and Human Decision Processes*, *122*, 36–52. https://doi.org/10.1016/j.obhdp.2013.04.001

Bahrami, B., Didino, D., Frith, C., Butterworth, B., & Rees, G. (2013). Collective enumeration. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 338–347. https://doi.org/10.1037/a0029717

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*, 3–8. https://doi.org/10.1037/a0025708

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 of the Royal Society B*, *367*, 1350–1365. https://doi.org/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*, 1081–1085. https://doi.org/10.1126/science.1185718

Bang, D., Aitchison, L., Moran, R., Castanon, S. H., Rafiee, B., Mahmoodi, A., … Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behavior*, *1*, 117. https://doi.org/10.1038/s41562-017-0117

Bang, D., Fusaroli, R., Tylen, K., Olsen, K., Latham, P., Lau, J., … Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and Cognition*, *26*, 13–23. https://doi.org/10.1016/j.concog.2014.02.002

Barnes, A. E., Nelson, T. O., Dunlosky, J., Mazzoni, G., & Narens, L. (1999). An integrative system of metamemory components involved in retrieval. *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application*, 287–313.

Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language*, *67*, 59–67. https://doi.org/10.1016/j.jml.2012.04.002

Budescu, D., & Rantilla, A. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398. https://doi.org/10.1016/S0001-6918(00)00037-8

Budescu, D., Rantilla, A., Yu, H., & Karelitz, T. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*, 178–194. https://doi.org/10.1016/S0749-5978(02)00516-2

Budescu, D., & Yu, H. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153–177. https://doi.org/10.1002/bdm.547

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York: Routledge.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Crawford, J., & Stankov, L. (1998). *Individual differences in the realism of confidence judgments Overconfidence in measures of fluid and crystallized intelligence*. Unpublished manuscript.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Del Missier, F., Mäntylä, T., & Bruin, W. B. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making*, *25*(4), 331–351. https://doi.org/10.1002/bdm.731

Del Missier, F., Mäntylä, T., & Nilsson, L. (2015). Aging, memory, and decision making. In T. Hess, J. Strough, & C. Löckenhoff (Eds.), *Aging and decision making: Empirical and applied perspectives* (pp. 127–149). San Diego, CA: Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-417148-0.00007-8

Djulbegovic, B., Elqayam, S., Reljic, T., Hozo, I., Miladinovic, B., Tsalatsanis, A., … Cannon-Bowers, J. (2014). How do physicians decide to treat: An empirical evaluation of the threshold model. *BMC Medical Informatics and Decision Making*, *14*:47. https://doi.org/10.1186/1472-6947-14-47

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment*, *18*, 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Estes, R., & Hosseini, J. (1988). The gender gap on wall street: An empirical analysis of confidence in investment decision making. *The Journal of Psychology*, *122*(6), 173–181.

Fetsch, C. R., Kiani, R., & Shadlen, M. N. (2014). Predicting the accuracy of a decision: A neural mechanism of confidence. *Cold Spring Harbor Symposia on Quantitative Biology*, *79*, 185–197. https://doi.org/10.1101/sqb.2014.79.024893

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 552–564. https://doi.org/10.1037/0096-1523.3.4.552

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98* (4), 506–528. https://doi.org/10.1037/0033-295X.98.4.506

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY, US: McGraw-Hill.

Harrison, D. A., Mohammed, S., McGrath, J. E., Florey, A. T., & Vanderstoep, S. W. (2003). Time matters in team performance: Effects of member familiarity, entrainment, and task discontinuity on speed and quality. *Personnel Psychology*, *56*(3), 633–669. https://doi.org/10.1111/j.1744-6570.2003.tb00753.x

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*(3), 305–326. https://doi.org/10.1006/obhd.1995.1024

Henry, R. A. (1993). Group judgment accuracy: Reliability and validity of postdiscussion confidence judgments. *Organizational Behavior and Human Decision Processes*, *56*, 11–27. https://doi.org/10.1006/obhd.1993.1043

Hill, G. W. (1982). Group versus individual performance: Are N + 1 heads better than one? *Psychological Bulletin*, *91*(3), 517–539. https://doi.org/10.1037/0033-2909.91.3.517

Hinsz, V. B. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social Psychology*, *59*(4), 705–718. https://doi.org/10.1037/0022-3514.59.4.705

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: Capturing decision tendencies in a fictitious medical test. *Metacognition and Learning*, *9*, 25–49. https://doi.org/10.1007/s11409-013-9110-y

Jackson, S. A., Kleitman, S., Howie, P., & Stankov, L. (2016). Cognitive abilities, monitoring confidence and control thresholds explain individual differences in heuristics and biases. *Frontiers in Psychology*, *7*, 1559. https://doi.org/10.3389/fpsyg.2016.01559

Jackson, S. A., Kleitman, S., Stankov, L., & Howie, P. (2016). Decision pattern analysis as a general framework for studying individual differences in decision making: Decision pattern analysis. *Journal of Behavioral Decision Making*, *29*, 392–408. https://doi.org/10.1002/bdm.1887

Jackson, S. A., Kleitman, S., Stankov, L., & Howie, P. (2017). Individual differences in decision making depend on cognitive abilities, monitoring and control. *Journal of Behavioral Decision Making*, *30*(2), 209–223. https://doi.org/10.1002/bdm.1939

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, *57*(2), 226–246. https://doi.org/10.1006/obhd.1994.1013

Keith, T. Z. (2006). *Multiple regression and beyond*. London: Pearson Education, Inc.

Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273.

Kleitman, S. (2008). *Metacognition in the rationality debate*. Germany: VDM Verlag Dr. Müeller Inc., Publishers.

Kleitman, S., & Gibson, J. (2011). Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students. *Learning and Individual Differences*, *21*(6), 728–735. https://doi.org/10.1016/j.lindif.2011.08.003

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology*, *15*(3), 321–341. https://doi.org/10.1002/acp.705

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*, 161–173. https://doi.org/10.1016/j.lindif.2007.03.004

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, *25*(4), 987–995.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 945–959. https://doi.org/10.1037/0278-7393.34.4.945

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, *140*, 117–139. https://doi.org/10.1037/a0022171

Koriat, A. (2012a). When are two heads better than one and why? *Science*, *336*, 360–362. https://doi.org/10.1126/science.1216549

Koriat, A. (2012b). The self-consistency model of subjective confidence. *Psychological Review*, *119*, 80–113. https://doi.org/10.1037/a0025648

Koriat, A. (2015). When two heads are better than one and when they can be worse: The amplification hypothesis. *Journal of Experimental Psychology: General*, *144*, 934–950. https://doi.org/10.1037/xge0000092

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517. https://doi.org/10.1037/0033-295X.103.3.490

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Laughlin, P. R. (2011). *Group problem solving* (Course Book ed.). Princeton, N.J: Princeton University Press.

Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes*, *88*(2), 605–620. https://doi.org/10.1016/S0749-5978(02)00003-1

Lauriola, M., & Irwin, P. (2001). Personality traits and risky decision-making in a controlled experimental task: An exploratory study. *Personality and Individual Differences*, *31*(2), 215–226. https://doi.org/10.1016/S0191-8869(00)00130-6

Littlepage, G. E., Schmidt, G. W., Whisler, E. W., & Frost, A. G. (1995). An input-process-output analysis of influence and performance in problem-solving groups. *Journal of Personality and Social Psychology*, *69*(5), 877–889. https://doi.org/10.1037/0022-3514.69.5.877

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley Pub. Co.

Mahmoodi, A., Bang, D., Olsen, K., Zhao, Y. A., Shi, Z., Broberg, K., … Bahrami, B. (2015). Equality bias impairs collective decision-making across cultures. *Proceedings of the National Academy of Sciences*, *112* (12), 3835-3840. https://doi.org/10.1073/pnas.1421692112

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*, 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Massoni, S., & Roux, N. (2017). Optimal group decision: A matter of confidence calibration. *Journal of Mathematical Psychology*, *79*, 121–130. https://doi.org/10.1016/j.jmp.2017.04.001

Michaelsen, L. K., Watson, W. E., & Black, R. H. (1989). A realistic test of individual versus group consensus decision making. *Journal of Applied Psychology*, *74*(5), 834–839. https://doi.org/10.1037/0021-9010.74.5.834

Minson, J. A., & Mueller, J. S. (2012). The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psychological Science*, *23*(3), 219–224. https://doi.org/10.1177/0956797611429132

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109–133. https://doi.org/10.1037/0033-2909.95.1.109

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*(2), 102–116. https://doi.org/10.1037/0003-066X.51.2.102

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology*, *129*(3), 257–299. https://doi.org/10.1080/00221300209602099

Pansky, A., & Goldsmith, M. (2014). Metacognitive effects of initial question difficulty on subsequent memory performance. *Psychonomic Bulletin & Review*, *21*, 1255–1262. https://doi.org/10.3758/s13423-014-0597-2

Pansky, A., Goldsmith, M., Koriat, A., & Pearlman-Avnion, S. (2009). Memory accuracy in old age: Cognitive, metacognitive, and neurocognitive determinants. *European Journal of Cognitive Psychology*, *21*, 303–329. https://doi.org/10.1080/09541440802281183

Patalano, A. L., & LeClair, Z. (2011). The influence of group decision making on indecisiveness-related decisional confidence. *Judgment and Decision making*, *6*(2), 163–175.

Pauker, S. G., & Kassirer, J. P. (1980). The threshold approach to clinical decision making. *The New England Journal of Medicine*, *302*(20), 1109–1117. https://doi.org/10.1056/NEJM198005153022003

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition. *Journal of Experimental Psychology: General*, *145*(8), 949–965. https://doi.org/10.1037/xge0000180

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. https://doi.org/10.1037/a0019737

Powell, M., & Ansic, D. (1997). Gender differences in risk behavior in financial decision-making: An experimental analysis. *Journal of Economic Psychology*, *18*(6), 605–628. https://doi.org/10.1016/S0167-4870(97)00026-3

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719. https://doi.org/10.1037/a0033152

Raven, J. C. (1938–65). *Progressive matrices*. New York: The Psychological Corporation.

Rudolph, J., Niepela, C., Greiffa, S., Goldhammerb, F., & Kröner, S. (2017). Metacognitive confidence judgments and their link to complex problem solving. *Intelligence*, *63*, 1–8. https://doi.org/10.1016/j.intell.2017.04.005

Savadori, L., Van Swol, L. M., & Sniezek, J. A. (2001). Information sampling and confidence within groups and judge advisor systems. *Communication Research*, *28*(6), 737-771.

Schraw, G., Dunkle, M. E., Bendixen, L. D., & Roedel, T. D. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology*, *87*(3), 433–444. https://doi.org/10.1037/0022-0663.87.3.433

Schuldt, J. P., Chabris, C. F., Woolley, A. W., & Hackman, J. R. (2017). Confidence in dyadic decision making: The role of individual differences. *Journal of Behavioral Decision Making*, *30*(2), 168–180. https://doi.org/10.1002/bdm.1927

Sniezek, J., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*(1), 1–28. https://doi.org/10.1016/0749-5978(89)90055-1

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288–307. https://doi.org/10.1006/obhd.2000.2926

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, *65*(2), 117–137. https://doi.org/10.1006/obhd.1996.0011

Stankov, L. (1997). *Gf–Gc quickie test battery*. *Gf–Gc Quickie Test Battery*. Sydney, Australia: E-ntelligence Testing Products.

Stankov, L. (1999). Mining on the "no man&apos;s land" between intelligence and personality. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (p. 315–337). American Psychological Association. https://doi:10.1037/10315-014

Stankov, L. (2013). Non-cognitive predictors of intelligence and academic achievement: An important role of confidence. *Personality and Individual Differences*, *55*, 727–732. https://doi.org/10.1080/01443410.2013.814194

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, *21*, 971–986. https://doi.org/10.1016/S0191-8869(96)00130-4

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence*, *25*(2), 93–109. https://doi.org/10.1016/S0160-2896(97)90047-7

Stankov, L., Kleitman, S., & Jackson, S. A. (2014). Measures of the trait of confidence. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 158–189). London, UK: Academic Press.

Stankov, L., & Lee, J. (2014). Quest for the best non-cognitive predictor of academic achievement. *Educational Psychology*, *34*, 1–8. https://doi.org/10.1080/01443410.2013.858908

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences*, *22*(6), 747-758. https://doi.org/10.1016/j.lindif.2012.05.013

Stankov, L., Pallier, G., Danthiir, V., & Morony, S. (2012). Perceptual underconfidence: A conceptual illusion? *European Journal of Psychological Assessment*, *28*(3), 190–200. https://doi.org/10.1027/1015-5759/a000126

Sunstein, C. R., & Hastie, R. (2015). *Wiser: Getting beyond groupthink to make groups smarter*. Boston: Harvard Business Review Press.

Tindale, R. S. (1989). Group vs individual information processing: The effects of outcome feedback on decision making. *Organizational Behavior and Human Decision Processes*, *44*(3), 454–473. https://doi.org/10.1016/0749-5978(89)90019-8

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971. https://doi.org/10.1037/a0037099

Wahn, B., Czeszumski, A., & Konig, P. (2018). Performance similarities predict collective benefits in dyadic and triadic joint visual search. *PLoS ONE*, *13*(1):e0191179. https://doi.org/10.1371/journal.pone.0191179

Worthy, D. A., Gorlick, M. A., Pacheco, J. L., Schnyer, D. M., & Maddox, W. T. (2011). With age comes wisdom. Decision making in younger and older adults. *Psychological Science*, *22*(11), 1375–1385. https://doi.org/10.1177/0956797611420301

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, *69*, 237–249. https://doi.org/10.1006/obhd.1997.2685

Yaniv, I. (2011). Group diversity and decision quality: Amplification and attenuation of the framing effect. *International Journal of Forecasting*, *27*, 41–49. https://doi.org/10.1016/j.ijforecast.2010.05.009

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 558–563. https://doi.org/10.1037/a0014589

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.

Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology*, *33*(4), 345–366. https://doi.org/10.1006/jesp.1997.1326

How to cite this article: Blanchard MD, Jackson SA, Kleitman S. Collective decision making reduces metacognitive control and increases error rates, particularly for overconfident individuals. *J Behav Dec Making*. 2020;33: 348–375. https://doi.org/10.1002/bdm.2156

APPENDIX A

CALCULATION OF THE POST

Logistic regression returns a logit function that predicts the bet decisions for any given value of confidence. For each confidence rating, the logit equation provides an estimate of the probability (p) that a bet will be made (bet decision = 1). When $p \geq .50$, the model predicts a bet was placed (1), and when $p < .50$, the model predicts no bet (0). The POST was defined as the point where the logit equation was equal to .50, which represented the level of confidence at which a bet decision changed from not placed (0) to placed (1). Using the parameters obtained from the logistic model, the POST was calculated as −1 multiplied by the intercept (a), divided by the unstandardized regression coefficient (b).

$$-1 \times a \div b.$$

If there was no variance in confidence ratings or bet decisions, then we could not calculate a POST. These instances were rare. We could not calculate a POST for three individuals. The POST is unique from the two decision tendency variables (i.e., decisiveness and recklessness) that we have calculated. For example, one can have a low POST and be low on recklessness if their confidence ratings tend to be low and they tend to place bets when their judgments are correct. Alternatively, one can have a low POST and be high on recklessness if they tend to have low confidence ratings and they tend to place bets when their judgments are incorrect.

APPENDIX B

GAMMA CORRELATIONS BETWEEN ACCURACY AND CONFIDENCE

We calculated within-person gamma correlations, which have been used to measure association in ordered tables, to further investigate the relationship between accuracy and confidence across questions (Bang et al., 2014; Koriat, 2012a, Koriat, 2015; Nelson, 1984). To calculate within-person gamma correlations, the questions were first split into two groups based on their consensuality (CC/CW), and then the responses of individuals and dyads (at the group level) were separated into independent strings of accuracy and confidence values. A gamma correlation was then calculated between accuracy and confidence for each individual and dyad for CC and CW questions.

For CC questions, the average correlation was significantly positive for individuals, γ = .25, t(103) = 2.64, p < .01, and for dyads, γ = .30, t(51) = 2.20, p = .03. For CW questions, the average

TABLE C1 Real dyads vs. virtual dyads: Series of 2 × 2 analyses of variance of the difference between grouping (real dyads vs. virtual dyads) and question consensuality (Consensually Correct vs. Consensually Wrong) on accuracy, confidence, decisiveness, and recklessness

	Mean difference	F(1, 51)	p	η^2	$\eta_p{}^2$
Accuracy					
Grouping	−0.44	0.11	.74	—	—
Consensuality	27.67	148.29	<.001***	.54	.74
Grouping × Consensuality	—	4.33	.04*	.01	.08
Confidence					
Grouping	2.73	33.09	<.001***	.06	.39
Consensuality	3.82	56.30	<.001***	.11	.52
Grouping × Consensuality	—	1.66	.20	—	—
POST					
Grouping	4.03	15.90	<.001***	.03	.18
Consensuality	−0.76	2.41	.13	—	—
Grouping × Consensuality	—	0.12	.73	—	—
Decisiveness					
Grouping	−0.96	0.33	.57	—	—
Consensuality	12.75	55.01	<.001***	.10	.52
Grouping × Consensuality	—	0.10	.76	—	—
Recklessness					
Grouping	2.79	1.45	.23	—	—
Consensuality	0.24	0.01	.94	—	—
Grouping × Consensuality	—	0.02	.89	—	—

Note. Two dyads demonstrated no recklessness, so the degrees of freedom for this analysis were lower than the others (df = 1, 49). POSTs could not be estimated for a number of dyads, so the degrees of freedom were lower for this analysis: grouping (df = 1, 50) and consensuality (df = 1, 47).
Abbreviation: POST, Point Of Sufficient cerTainty.
$^{***}p$ < .001.$^{*}p$ < .05.

correlation was significantly negative for individuals, $\gamma = -.25$, $t(103) = -2.58$, $p = .01$, and for dyads, $\gamma = -.22$, $t(51) = -1.62$, $p = .11$; however, potentially due to the reduced sample size, it was not significant. Overall, this pattern is consistent with Koriat (2012a, 2015) and suggests that participants were not aware of the deceptive nature of CW questions.

APPENDIX C

VIRTUAL DYADS VERSUS REAL DYADS

Using a maximum confidence slating algorithm (Bang et al., 2014; Koriat, 2012a, 2015; Pescetelli et al., 2016), we created virtual dyads to identify the theoretically account of what should occur if groups rely on confidence to guide their judgments. This was done by selecting the general-knowledge responses (judgment, confidence rating, and bet decision) of the most confident individual in each dyad for all 30 questions. If both members responded with the same level of individual confidence, then we randomly selected one member's responses. This process was consistent with the real dyad dataset. We then calculated each outcome variable (i.e., accuracy, confidence, POST, decisiveness, and recklessness) for the virtual dyads.

We used a series of 2 × 2 repeated-measures ANOVAs to compare the outcome variables for real dyads with those of the virtual dyads. The independent variables were grouping (virtual dyads vs. real dyads) and question consensuality (CC vs. CW). Table C1 reveals the results of these analyses.

For accuracy, there was no difference between virtual dyads and real dyads. As expected, accuracy was significantly higher for CC than CW questions. The interaction, however, was significant, so we conducted simple effects tests to interpret. Accuracy was significantly higher for CC questions than CW questions for both real dyads, mean difference = 25.10, $t(51) = 9.58$, $p < .001$, $d = 2.04$, and virtual dyads, mean difference = 30.25, $t(51) = 11.84$, $p < .001$, $d = 2.27$. These results suggest that the difference on accuracy between CC and CW questions was more pronounced for virtual dyads than real dyads. It should be noted, however, that we did not observe an accuracy difference between virtual dyads and real dyads overall, for CC questions, mean difference = −2.13, $t(51) = -1.18$, $p = .24$, or CW questions, mean difference = 3.01, $t(51) = 1.65$, $p = .10$.

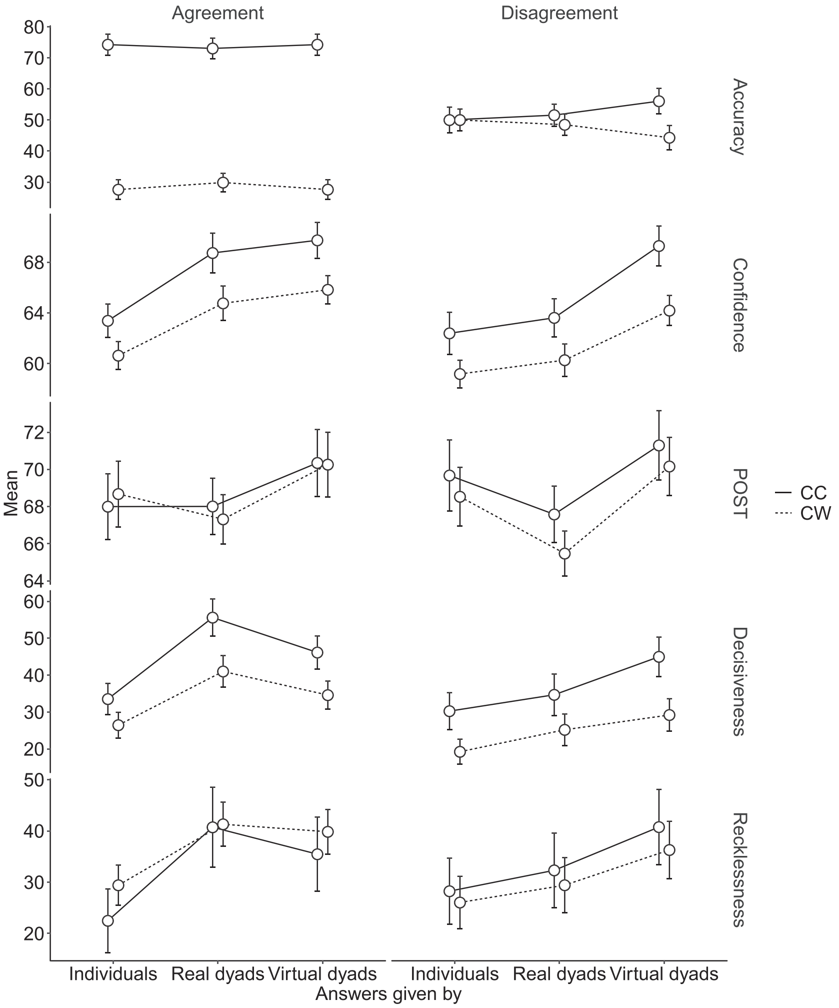

FIGURE D1 Mean score for general-knowledge test variables depending on grouping (Individuals vs. Real Dyads vs. Virtual Dyads), consensus (Agreement vs. Disagreement), and question consensuality (CC vs. CW). Error bars represent 95% confidence intervals

For confidence, virtual dyads were significantly more confident than real dyads, and confidence was significantly higher for CC questions than CW questions. The interaction was not significant. These findings suggest that the confidence of real dyads increased but not to the level of the most confident member.

For the POST, virtual dyads had significantly higher POSTs than real dyads, regardless of question consensuality. Furthermore, POSTs did not differ for CC or CW questions, and the interaction was not significant. This finding suggests that the POSTs of real dyads did not increase to the level of the most confident member.

For decisiveness, there was no difference between virtual dyads and real dyads, and decisiveness was significantly greater for CC than CW questions. The interaction was not significant.

For recklessness, there was no difference between virtual dyads and real dyads or CC and CW questions. The interaction was also not significant. These results indicate that there was no difference between virtual dyads and real dyads in the overall proportion of bets lost (false alarm rate).

APPENDIX D

EFFECTS OF AGREEMENT ON THE DYADIC RESPONSE AND DECISION

When responding individually, dyad members entered the same true–false response 55% of the time, on average, for the General-knowledge Test. In 96% of these agreement trials, the dyadic response was the same as the individual response. On average, dyad members disagreed 45% of the time. Both dyad members entered the same level of confidence in 25% of these trials; these instances were removed. Of the remaining trials, 67% of the true–false responses endorsed by the dyad were the individual responses that had been associated with higher confidence. At the item level, the most confident member's response was endorsed by the dyad for 69% of CC items and 66% of CW items. Furthermore, when the high confidence member's true–false response was endorsed by the dyad, their bet decision was endorsed on 74% of these trials. Thus, dyads endorsed the response and the bet decision of the most confident member on the majority of trials.

We ran a series of three-way ANOVAs to include consensus as an additional variable to grouping and question consensuality (as reported in Section 3). Figure D1 shows the mean accuracy, confidence, decisiveness (overall proportion of bets), and recklessness (false alarm rate) for agreement trials and disagreement trials on CC and CW questions for the General-knowledge Test when answered by individuals and real dyads. Descriptive statistics are shown in Table D1. With the exception of accuracy, reliability estimates tended to be acceptable at the overall level. Likely due to creating small subsets of data, when questions were split into CC and CW, the reliability estimates tended to be poor. We have included consensuality in the proceeding analyses but would be hesitant to generalize the findings related to this variable.

D.1 | Individuals versus real dyads

We examined the decisions of the real dyads compared with when the same participants acted individually. The independent variables were grouping (individuals vs. real dyads), consensus (agreement vs. disagreement), and question consensuality (CC vs. CW). Table D2 reveals the results of this series of ANOVAs.

For accuracy, there was no difference between individuals and real dyads or agreement and disagreement trials. As expected, accuracy was greater for CC questions than CW questions. The interaction between consensus and question consensuality was significant; thus, to interpret this effect, we conducted simple effects tests. For agreement trials, accuracy was significantly higher for CC than CW questions, mean difference = 44.92, $t(51) = 14.38$, $p < .001$, $d = 2.74$. For disagreement trials, unexpectedly, there was no difference on accuracy between CC and CW questions, mean difference = 1.54, $t(51) = 0.81$, $p = .42$, indicating that disagreement may mitigate the harmful influence of CW questions on judgment accuracy. None of the other interactions were significant.

Confidence was significantly higher for real dyads than individuals, agreement trials than disagreement trials, and CC than CW questions. To interpret the significant interaction between grouping and consensus, we conducted simple effects tests. For agreement trials, confidence was significantly higher for real dyads than individuals, mean difference = −4.74, $t(51) = -10.80$, $p < .001$, $d = 0.97$. For disagreement trials, confidence was also significantly higher for real dyads than individuals, mean difference = −1.16, $t(51) = -2.81$, $p < .01$, $d = 0.24$. This pattern of results indicates that confidence was higher for real dyads than individuals for both levels of consensus and this effect was more pronounced for agreement trials than disagreement trials. None of the other interactions were significant.

The POST was significantly lower for real dyads than individuals, regardless of question consensuality or consensus (agree vs. disagree). There was no difference on the POST between agreement and disagreement trials or CC and CW questions. Additionally, none of the interactions were significant.

Decisiveness was significantly higher for real dyads than individuals, agreement than disagreement trials, and CC than CW questions. We also observed a significant interaction between type of grouping and consensus and another significant interaction between all three independent variables. To interpret the three-way interaction, we conducted simple effects tests. On agreement trials, decisiveness was significantly higher for dyads than individuals for CC questions, mean difference = −22.10, $t(51) = -9.16$, $p < .001$, $d = 0.98$, and CW questions, mean difference = −14.56, $t(51) = -7.00$, $p < .001$, $d = 0.76$. On disagreement trials, decisiveness was significantly higher for dyads than individuals for CW questions, mean difference = −5.96, $t(51) = -2.67$, $p = .01$, $d = 0.32$, but there was no difference for CC questions, mean difference = −4.43, $t(51) = -1.31$, $p = .20$. These results reveal that decisiveness was greater for dyads than individuals and this difference was more pronounced for CC questions on agreement trials but CW questions on disagreement trials. None of the other interactions were significant.

TABLE D1 Descriptive statistics for individuals and real dyads for agreement and disagreement trials

	Individual						Real dyad					
	Agree			Disagree			Agree			Disagree		
Variable	Mean	*SD*	RE	Mean	*SD*	RE	Mean	*SD*	RE	Mean	*SD*	RE
Accuracy (% correct)												
Total	50.93	28.91	.06	50.00	19.96	.00	51.46	27.09	−.24	50.02	18.17	.10
CC	74.26	17.54	−.26	50.00	21.54	.16	73.05	17.20	−.36	51.56	18.34	.19
CW	27.60	16.53	.22	50.00	18.35	−.24	29.87	15.42	.11	48.48	17.95	−.24
Confidence (mean %)												
Total	62.00	6.47	.67	60.78	7.44	.68	66.74	7.87	.79	61.94	7.43	.61
CC	63.38	6.87	.65	62.39	8.62	.71	68.73	8.13	.72	63.61	7.82	.38
CW	60.63	5.75	.55	59.18	5.64	.59	64.76	7.10	.75	60.27	6.63	.80
POST (mean %)												
Total	68.34	9.17	.48	69.09	9.10	.65	67.65	7.39	.61	66.48	7.17	.38
CC	67.99	9.19	.41	69.67	9.96	.31	68.00	7.90	.38	67.58	7.92	−.41
CW	68.67	9.19	.57	68.53	8.23	.63	67.31	6.88	−.02	65.47	6.30	.63
Decisiveness (% of bets placed)												
Total	29.98	20.41	.64	24.73	22.86	.59	48.31	25.41	.69	29.93	26.35	.65
CC	33.52	22.03	.49	30.24	26.01	.45	55.62	26.22	.33	34.67	29.27	.31
CW	26.44	18.06	.45	19.23	17.69	.40	41.00	22.41	.46	25.18	22.22	.70
Recklessness (% of bets lost)												
Total	26.27	26.61	.48	27.14	30.14	.34	41.04	31.98	.52	30.87	33.20	.42
CC	22.48	32.26	.26	28.25	33.52	.57	40.72	40.60	.62	32.32	37.79	.54
CW	29.41	20.45	.26	26.04	26.52	.20	41.30	22.47	.25	29.42	27.98	.33

Note. All internal reliability was estimated by correlating the odd and even questions corrected by the Spearman–Brown prophecy formula (Guilford, 1954; Jackson et al., 2017; Stankov & Crawford, 1996).
Abbreviations: CC, Consensually Correct; CW, Consensually Wrong; POST, Point Of Sufficient cerTainty.

For recklessness, dyads were significantly more reckless than individuals, and there was no significant main effect for consensus or question consensuality. To interpret the significant interaction between type of grouping and consensus, we conducted simple effects tests. Recklessness was significantly higher for dyads than individuals on agreement trials, mean difference = −13.80, *t*(51) = −4.90, *p* < .001, *d* = 0.66, but there was no difference on disagreement trials, mean difference = −4.15, *t*(51) = −1.30, *p* = .20. None of the other interactions were significant.

D.2 | Real dyads versus virtual dyads

Next, we compared the observed results for the real dyads with the virtual dyads, which represent the theoretical account of what would occur if groups relied on confidence. The independent variables were grouping (virtual dyads vs. real dyads), consensus (agreement vs. disagreement), and question consensuality (CC vs. CW). Table D3 reveals the results of this series of ANOVAs.

For accuracy, there was no difference between real dyads and virtual dyads or agreement and disagreement trials. As expected, accuracy was significantly greater for CC questions than CW questions. There was, however, a significant interaction between grouping and question consensuality; thus, we ran simple effects tests to interpret it. Accuracy was significantly higher for CC than CW questions for real dyads, mean difference = 23.13, *t*(51) = 8.52, *p* < .001, *d* = 1.80, and for virtual dyads, mean difference = 29.21, *t*(51) = 10.81, *p* < .001, *d* = 2.09. These results suggest that the difference on accuracy between CC and CW questions was more pronounced for virtual dyads than real dyads. It should be noted, however, that we did not observe an accuracy difference between virtual dyads and real dyads overall, for CC questions, mean difference = −2.86, *t*(51) = −1.48, *p* = .15, or CW questions, mean difference = 3.22, *t*(51) = 1.61, *p* = .11. We also conducted simple effects tests to interpret the significant interaction between consensus and question consensuality. For agreement trials, accuracy was significantly higher for CC than CW questions, mean difference = 44.92, *t*(51) = 14.38, *p* < .001, *d* = 2.74. For disagreement trials, there was no difference in accuracy between CC and CW questions, mean difference = 7.42, *t*(51) = 2.46, *p* = .02. None of the other interactions were significant.

For confidence, significant main effects were observed for grouping, consensus, and question consensuality. We also found a significant interaction between grouping and consensus and a significant three-way interaction between all independent variables. We conducted simple effects tests to interpret these effects. For agreement

TABLE D2 Individuals versus real dyads: Series of three-way analyses of variance of the difference between grouping (real dyads vs. virtual dyads), consensus (agreement vs. disagreement), and question consensuality (Consensually Correct vs. Consensually Wrong) on accuracy, confidence, decisiveness, and recklessness

	Mean difference	*F*(1, 51)	*p*	η^2	η_p^2
Accuracy					
Grouping	−0.27	0.08	.79	—	—
Consensus	1.19	0.38	.54	—	—
Consensuality	23.23	133.88	<.001***	.38	.72
Grouping × Consensus	—	0.10	.76	—	—
Grouping × Consensuality	—	0.01	.92	—	—
Consensus × Consensuality	—	178.77	<.001***	.35	.78
Grouping × Consensus × Consensuality	—	2.58	.12	—	—
Confidence					
Grouping	−2.95	67.30	<.001***	.07	.57
Consensus	3.01	35.60	<.001***	.07	.41
Consensuality	3.32	65.88	<.001***	.09	.56
Grouping × Consensus	—	61.66	<.001***	.03	.55
Grouping × Consensuality	—	3.67	.06†	—	—
Consensus × Consensuality	—	0.01	.93	—	—
Grouping × Consensus × Consensuality	—	3.10	.08†	—	—
POST					
Grouping	1.53	5.96	.02*	.01	.03
Consensus	0.19	1.39	.25	—	—
Consensuality	0.75	0.22	.64	—	—
Grouping × Consensus	—	0.82	.37	—	—
Grouping × Consensuality	—	0.31	.58	—	—
Consensus × Consensuality	—	1.76	.19	—	—
Grouping × Consensus × Consensuality	—	0.01	.91	—	—
Decisiveness					
Grouping	−11.76	55.37	<.001***	.07	.52
Consensus	11.82	40.96	<.001***	.07	.45
Consensuality	10.55	39.17	<.001***	.06	.43
Grouping × Consensus	—	22.29	<.001***	.02	.30
Grouping × Consensuality	—	1.81	.18	—	—
Consensus × Consensuality	—	0.02	.88	—	—
Grouping × Consensus × Consensuality	—	5.40	<.01**	.002	.10
Recklessness					
Grouping	−8.97	23.95	<.001***	.03	.41
Consensus	−1.06	2.29	.14	—	—
Consensuality	−1.06	0.19	.66	—	—
Grouping × Consensus	—	5.85	.02*	.03	.41
Grouping × Consensuality	—	0.08	.78	—	—
Consensus × Consensuality	—	0.51	.48	—	—
Grouping × Consensus × Consensuality	—	2.78	.10	—	—

Note. Two dyads demonstrated no recklessness, so the degrees of freedom for this analysis were lower than the others (*df* = 1, 49). POSTs could not be estimated for all individuals or dyads, so the degrees of freedom were lower for this analysis: grouping (*df* = 1, 46), consensus (*df* = 1, 43), and consensuality (*df* = 1, 43).

Abbreviation: POST, Point Of Sufficient cerTainty.

****p* < .001.

***p* < .01. **p* < .05. †*p* < .10.

TABLE D3 Individuals versus real dyads: Series of three-way analyses of variance of the difference between grouping (real dyads vs. virtual dyads), consensus (agreement vs. disagreement), and question consensuality (Consensually Correct vs. Consensually Wrong) on accuracy, confidence, decisiveness, and recklessness

	Mean difference	$F(1, 51)$	p	η^2	η_p^2
Accuracy					
Grouping	–0.18	0.02	.90	–	–
Consensus	1.09	0.27	.61	–	–
Consensuality	26.17	128.00	$<.001^{***}$	.34	.72
Grouping × Consensus	–	4.65	$.04^{*}$	.01	.08
Grouping × Consensuality	–	0.07	.79	–	–
Consensus × Consensuality	–	86.60	$<.001^{***}$	.21	.63
Grouping × Consensus × Consensuality	–	0.96	.33	–	–
Confidence					
Grouping	–2.91	36.83	$<.001^{***}$	.05	.42
Consensus	2.92	25.34	$<.001^{***}$	.05	.33
Consensuality	4.08	60.37	$<.001^{***}$	.09	.54
Grouping × Consensus	–	45.30	$<.001^{***}$	.02	.47
Grouping × Consensuality	–	2.73	.10	–	–
Consensus × Consensuality	–	0.05	.82	–	–
Grouping × Consensus × Consensuality	–	5.68	.03	.001	.10
POST					
Grouping	3.34	11.01	$<.01^{**}$	.03	.09
Consensus	–0.06	1.32	.26	–	–
Consensuality	0.70	0.13	.72	–	–
Grouping × Consensus	–	1.08	.30	–	–
Grouping × Consensuality	–	0.40	.53	–	–
Consensus × Consensuality	–	1.61	.21	–	–
Grouping × Consensus × Consensuality	–	0.00	.97	–	–
Decisiveness					
Grouping	0.39	0.05	.83	–	–
Consensus	10.83	21.87	$<.001^{***}$	. 05	.30
Consensuality	12.86	45.96	$<.001^{***}$	.07	.47
Grouping × Consensus	–	30.04	$<.001^{***}$	.02	.37
Grouping × Consensuality	–	0.45	.50	–	–
Consensus × Consensuality	–	0.008	.93	–	–
Grouping × Consensus × Consensuality	–	4.10	.048	.002	.07
Recklessness					
Grouping	2.03	0.90	.35	–	–
Consensus	3.85	0.85	.36	–	–
Consensuality	0.58	0.03	.86	–	–
Grouping × Consensus	–	3.06	$.09^{\dagger}$	–	–
Grouping × Consensuality	–	0.01	.95	–	–
Consensus × Consensuality	–	0.93	.34	–	–
Grouping × Consensus × Consensuality	–	0.35	.56	–	–

Note. Two dyads demonstrated no recklessness, so the degrees of freedom for this analysis were lower than the others (*df* = 1, 49). POSTs could not be estimated for all dyads, so the degrees of freedom were lower for this analysis: grouping (*df* = 1, 46), consensus (*df* = 1, 43), and consensuality (*df* = 1, 43).
Abbreviation: POST, Point Of Sufficient cerTainty.
$^{***}p < .001.$ $^{**}p < .01.$ $^{*}p < .05.$ $^{\dagger}p < .10.$

trials, there was no difference on confidence between real dyads and virtual dyads for CC questions, mean difference = −1.00, *t*(51) = −1.70, *p* = .10, or CW questions, mean difference = −1.07, *t*(51) = −1.67, *p* = .10. For disagreement trials, confidence was significantly greater for virtual dyads than real dyads for CC questions, mean difference = −7.49, *t*(51) = −5.67, *p* < .001, *d* = 0.80, and CW questions, mean difference = −3.91, *t*(51) = −7.04, *p* < .001, *d* = 0.68. This pattern indicates that real dyads increased in confidence to the level of the most confident member (virtual dyads) on agreement trials, but not disagreement trials, and the difference between real dyads and virtual dyads was more pronounced for CC questions than CW questions. It is worth noting that real dyads were more confident than individuals for disagreement trials. None of the other interactions were significant.

The POST was significantly lower for real dyads than virtual dyads. None of the other main effects or interactions were significant.

For decisiveness, there was no difference between real dyads and virtual dyads. We found significant main effects for consensus and question consensuality. We did, however, observe a significant interaction between type of grouping and consensus and a significant interaction between all three independent variables. To interpret these effects, we ran a series of simple effects tests. On agreement trials, decisiveness was significantly higher for real dyads than virtual dyads for CC questions, mean difference = 9.47, *t*(51) = 3.92, *p* < .001, *d* = 0.38, and CW questions, mean difference = 6.41, *t* (51) = 2.80, *p* < .01, *d* = 0.30. On disagreement trials, decisiveness was significantly higher for virtual dyads than real dyads for CC questions, mean difference = −10.31, *t*(51) = −2.80, *p* < .01, *d* = 0.36, but there was no difference for CW questions, mean difference = −4.02, *t* (51) = −1.58, *p* = .12. None of the other interactions were significant. These results reveal that real dyads increased on decisiveness more than the virtual dyads for agreement trials. For disagreement trials, real dyads did not increase to the level of the virtual dyads for CC questions but did for CW questions.

For recklessness, there was no effect of grouping, consensus, or question consensuality. In addition, none of the interactions were significant. This pattern suggests that there was no difference between real dyads and virtual dyads on recklessness.

The results of this series of analyses reveal that the responses of real dyads were consistent with those of the virtual dyads and the confidence theory: Group members share and use each other's subjective confidence to guide their decisions. It is worth noting that the effects of working in a dyad were more pronounced when members initially agreed on a group response than when they initially disagreed.

APPENDIX E

PEARSON CORRELATIONS BETWEEN REGRESSION VARIABLES

The correlations between variables used in the regression analyses are reported in Table E1. All correlations were based on a dataset using dyads as the unit of analysis instead of using individual responses (*n* = 52). This is consistent with prior research (Bahrami et al., 2010; Koriat, 2012a, 2015; Schuldt et al., 2017).

Of the variables considered, all dyad–individual difference variables except accuracy correlated significantly and positively with each other (*r* = .47 to.84, *p* < .01) and with trait-confidence (*r* = .29 to.31, *p* < .05) and bias (*r* = .37 to.48, *p* < .05). The dyad–individual difference variables for decisiveness (overall proportion of bets) and recklessness (false alarm rate) also correlated significantly and negatively with Neuroticism (*r* = −.29 and − .31, *p* < .05). There was a similar pattern of correlations between Extraversion and the dyad–individual difference variables for confidence (*r* = .27, *p* = .05) and decisiveness (*r* = .25, *p* = .07). Although these correlations did not reach statistical significance, they appear to be meaningful (Stankov, 2013; Stankov & Lee, 2014), and the pattern appears to be nonrandom (correlated with more than one dyad–individual difference variable). No other significant and meaningful correlations were observed between the dyad–individual difference variables and the control variables. Thus, they were omitted from further analyses as they were deemed to hinder the power without introducing any useful control (Keith, 2006). Therefore, trait-confidence, Extraversion, and Neuroticism only were included in the regression analyses.

TABLE E1 Pearson correlations between regression variables

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Dyad–individual difference scores																
1. Accuracy	–.01	–.07	–.16	.22	–.08	–.03	–.05	–.11	.10	.01	.00	–.14	–.18	–.22	–.02	–.02
2. Confidence	1	.35	.58***	.44**	.29*	.48***	–.06	.05	–.22	.27†	–.10	–.18	.11	.20	.02	.01
3. POST		1	–.14	–.10	.21	.23	–.08	–.02	–.20	.22	.01	.05	.08	–.07	–.08	–.17
4. Decisiveness			1	.79***	.31*	.37**	.02	.15	.13	.25†	.09	–.29*	.06	.22	.12	.05
5. Recklessness				1	.29*	.39**	.01	.06	.13	.10	.07	–.31*	.02	.16	.17	–.10
Individual difference variables																
6. Trait-confidence					1	.30*	.62***	.12	–.08	–.01	.33*	–.10	.21	.25†	.29*	–.02
7. Bias						1	–.42**	.01	.01	.29*	–.02	–.18	.05	.21	.11	.16
8. Cognitive ability							1	.11	–.11	–.17	.41**	.02	.14	.07	.10	–.15
9. Agreeableness								1	.17	.24†	.18	–.14	.06	–.01	.14	.17
10. Conscientiousness									1	.04	–.09	.05	–.23†	–.17	–.10	.52***
11. Extraversion										1	–.07	–.23†	.09	–.09	–.13	.09
12. Intellect											1	–.10	–.17	.05	.04	–.07
13. Neuroticism												1	–.02	–.08	–.31*	–.26†
14. Gender													1	.02	.19	.00
15. English ability														1	.22	–.19
16. Member relationship															1	.08
17. Age																1

Note. Correlations were calculated using the dyad-level dataset.

Abbreviation: POST, Point Of Sufficient cerTainty.

$^{*}p < .05$.

$^{**}p < .01$.

$^{***}p < .001$.

$^{\dagger}p < .10$.