Journal of Behavioral Decision Making, *J. Behav. Dec. Making*, **30**: 168–180 (2017)
Published online 29 December 2015 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/bdm.1927

Confidence in Dyadic Decision Making: The Role of Individual Differences

JONATHON P. SCHULDT[1*,†], CHRISTOPHER F. CHABRIS[2,†], ANITA WILLIAMS WOOLLEY[3] and J. RICHARD HACKMAN[4]
[1]*Department of Communication, Cornell University, Ithaca, NY, USA*
[2]*Department of Psychology, Union College, Schenectady, NY, USA*
[3]*Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA*
[4]*Department of Psychology, Harvard University, Cambridge, MA, USA*

ABSTRACT

Groups typically express more confidence than individuals, yet how individual-level confidence combines during collaborative decision tasks is not well understood. We prescreened 686 community members using a novel confidence measure (a true/false trivia test) intentionally designed to be difficult (accuracy rates were not significantly better than chance) and randomly assigned 72 individuals to collaborate on a matched version of the same test in dyads composed of two low-confidence individuals, two high-confidence individuals, or one of each ("mixed"). Consistent with past research, we found that the confidence expressed by dyads was higher than the confidence expressed by individuals; importantly, however, this pattern varied markedly by dyad type, with low-confidence dyads showing the largest increase, mixed dyads showing a moderate increase, and high-confidence dyads showing no increase—despite the fact that all dyads showed similarly low accuracy (about 55%). These results highlight the conditions under which groups express greater confidence than individuals and offer insights for the composition of collaborative decision-making teams. Copyright © 2015 John Wiley & Sons, Ltd.

KEY WORDS confidence; overconfidence; collaborative teams; dyads; group decision making

What factors affect a group's level of confidence in its decisions? This question has long been a topic of inquiry among scholars of group dynamics, from the seminal work on risky shift within the group polarization literature (e.g., Stoner, 1961; Myers & Lamm, 1975) to more recent efforts to illuminate the conditions under which "two heads are better than one" (Bahrami et al., 2010; Koriat, 2012; Turner & Pratkanis, 1998). Accordingly, much of this research has focused on relating a group's confidence to the accuracy of its decisions, with the familiar finding that groups routinely fail to perform as well as their best individuals because of a variety of pitfalls (or "process losses") that can plague group settings (e.g., groupthink; Janis, 1972; see Kerr & Tindale, 2004, for a review).

Although this focus on the relationship between group confidence and accuracy has yielded important insights, consequential decisions are routinely based on the confidence expressed by groups long before the accuracy of those decisions is known, highlighting a need to better understand the factors that shape group confidence itself. For instance, consider the George W. Bush Administration's famously high confidence that Iraq possessed weapons of mass destruction in the lead up to the 2003 U.S. invasion. According to journalist Bob Woodward (2004), CIA Director George Tenet expressed his agency's confidence to the president by declaring, "It's a slam dunk case! … Don't worry, it's a slam dunk!" In turn, White House Press Secretary Ari Fleisher told the public,

> … we have high confidence that they have weapons of mass destruction. This is what this war was about and it is about. And we have high confidence it will be found (Fleischer, 2003).

In this vein, the present research focuses primarily on confidence itself, as opposed to the relationship between confidence and accuracy (or confidence realism; Adams & Adams, 1961) in exploring how confidence expressed by decision makers acting individually may shape the confidence expressed by groups they comprise. In doing so, we draw on research to suggest that confidence is equally or perhaps more reliable over time and across domains than is overconfidence, and that confidence is a suitable construct for the kind of individual difference research we pursue here.

We first review work on the reliability and trait-like nature of individuals' confidence judgments before turning to the present work: an experimental study in which participants were prescreened to assess their expressions of confidence on a general-knowledge test when working individually (as relatively low, medium, or high) and were later assigned to complete a similar test as a member of a collaborative dyad. Depending on experimental condition, dyads were composed of two low-confidence individuals (low condition), two high-confidence individuals (high condition), or one of each (mixed condition), allowing us to explore whether and how the confidence expressed by two people making joint decisions varies as a function of their individual-level confidence expressions.

Individual-level confidence and overconfidence

The bulk of research on confidence comes from studies that seek to relate an individual's subjective confidence in his or her decisions to a measure of task performance or accuracy (e.g., Lichtenstein, Fischhoff, & Phillips, 1982). The relationship between subjective confidence and some objective outcome measure, or *confidence realism* (Adams & Adams, 1961), is typically represented by subtracting overall performance from overall confidence, such that positive numbers

*Correspondence to: Jonathon P. Schuldt, Department of Communication, Cornell University, Ithaca, NY, USA. E-mail: jps56@cornell.edu
†These authors contributed equally to the work.

signal overconfidence and negative numbers signal underconfidence (Lichtenstein et al., 1982; Yates, 1990). Research in this domain reliably finds that individuals are unjustifiably confident in their decisions, exhibiting marked overconfidence in a wide array of domains including tests of general knowledge (Fischhoff, Slovic, & Lichtenstein, 1977; Koriat, Lichtenstein, & Fischhoff, 1980), cognitive ability (Pallier et al., 2002), and various social predictions (Dunning, Griffin, Milojkovic, & Ross, 1990). Although overconfidence itself is widely observed, its magnitude has been shown to vary widely by task difficulty and domain, with substantial overconfidence reported in low-accuracy tasks and less overconfidence reported on high-accuracy tasks. Known as the *hard–easy effect* (Baranski & Petrusic, 1994; Lichtenstein & Fischhoff, 1977, 1980), this apparent dependence of overconfidence on task difficulty has prompted discussion as to whether the phenomenon results from a real and pervasive cognitive bias or whether it is an artifact of the difficult, artificial, and potentially misleading nature of the tasks that are commonly posed to participants in laboratory settings (Gigerenzer, Hoffrage, & Kleinbölting, 1991).

Individual differences in confidence versus overconfidence

Although a handful of studies report evidence that is consistent with a general overconfidence trait (e.g., Buratti, Allwood, & Johansson, 2014; Jonsson & Allwood, 2003; Stankov & Crawford, 1996; West & Stanovich, 1997), the weight of the evidence appears to suggest that confidence itself, independent of considerations of accuracy, may be the more reliable construct—a notion with deep theoretical roots (Henmon, 1911; Johnson, 1939) and modern empirical support. Support for a domain-general confidence trait has emerged from studies reporting robust intercorrelations for individuals' confidence, but not necessarily for accuracy and overconfidence, across a range of tasks (Blais, Thompson, & Baranski, 2005; Kleitman & Stankov, 2001, 2007; Pallier et al., 2002; Schraw, 1997; Stankov, 1998). For instance, Bornstein and Zickafoose (1999) examined the relationships among participants' confidence, accuracy, and overconfidence across eyewitness memory and general-knowledge domains and found that confidence exhibited the strongest correlation reported in the entire study ($r = .49$). In other work, Blais et al. (2005) examined confidence estimates for forced-choice tasks spanning three domains (vocabulary, general-knowledge test, and a perceptual line length task) and found that, for each pair of tests, the confidence correlation exceeded the overconfidence correlation.[1] From a measurement perspective, observing stronger correlations for confidence than overconfidence is not surprising given that overconfidence is derived from two empirical measurements (i.e., confidence and accuracy) and therefore incorporates two sources of measurement error that combine to decrease the reliability of the resulting measure.[2]

Confidence expressions of individuals versus groups

Especially relevant to the present research is the question of whether confidence is similarly stable across the individual and group contexts. On this point, numerous studies suggest that groups typically express more confidence in their assessments than do individuals working alone. Allwood and Björhag (1990) recruited participants to complete a general-knowledge test working either individually or as a member of a collaborative dyad. For each response, participants were instructed to assign a confidence estimate on a scale from 50% to 100%. Results showed that dyads expressed significantly greater confidence than individuals but showed no commensurate gain in accuracy, an observation the authors highlight as indicative of risky shift (Stoner, 1961). It should be noted, however, that the between-subjects design of that study did not allow the researchers to compare the confidence levels expressed by the *same* people across the individual and group settings. In this vein, a subsequent study by Allwood and Granhag (1996) had individuals answer and provide item-level confidence expressions for 30 general-knowledge questions, before doing the same for a different set of 30 questions while providing one argument in support of each chosen answer. Later on, the participants were divided into dyads to collaboratively answer and rate their joint confidence in the second set of questions once more. Results showed that group confidence exceeded individual confidence in both of the individual conditions.[3] In a study explicitly examining the role of individual-level confidence in the confidence expressed by groups, Sniezek and Henry (1989) (see also Sniezek, 1992) used a common confidence measure in which individuals set 99% confidence intervals around frequency estimates for 15 causes of death (e.g., Haran, Moore, & Morewedge, 2010). Immediately thereafter, the same individuals completed the same task, with identical content, as members of randomly assigned triads. Results again showed that groups expressed significantly greater confidence than individuals; interestingly, groups were also more accurate and less overconfident, setting narrower confidence intervals that more frequently contained the correct point value.

Although related to the present work, our study departs from these past studies in notable ways. First, by directly assessing both individual-level and dyadic confidence on similar (but not identical) tasks, our design facilitates strong inferences about the influence of trait confidence and decision context on confidence expressions. Also, in contrast to expressions of confidence made by individuals about group

[1]Vocabulary and general knowledge, $r = .81$ for confidence (.62 for overconfidence); vocabulary and line length, $r = .30$ (.20); general knowledge and line length, $r = .38$ (.37; only vocabulary and general knowledge significantly correlated in terms of accuracy, $r = .57$)

[2]As Buratti et al. (2014) discuss, inconsistent results regarding the relative stability of confidence and over/underconfidence across studies may be attributable to different methods (e.g., the nature of the decision task) or different analytic techniques. For instance, although they reported stronger correlations for confidence than for overconfidence in a memory task across three timepoints, more advanced multilevel modeling revealed relatively little intra-individual stability for confidence as compared with overconfidence.

[3]However, the difference was statistically significant in comparison with the argument condition only.

DOI: 10.1002/bdm

decisions, we examine the reported confidence reached *collaboratively* by dyads.[4] Moreover, in addition to examining homogenous dyads (e.g., composed exclusively of low-confidence or high-confidence individuals), we examine heterogeneous dyads, allowing insight into the relative influence of individuals with different confidence tendencies over dyadic confidence decisions. Last, in light of the well-established tendency for individuals and groups to be overconfident in their decisions, our true/false questions were intentionally selected to engender low-accuracy rates (close to 50%) with the goal of inducing wider variation in confidence scores and increasing the statistical power of our design. At the same time, featuring questions with objectively correct answers allows us to explore the influence of individual-level confidence not only on the confidence expressions of dyads but also on dyadic accuracy and overconfidence.

The present work

The goal of the present research is to investigate how dyadic confidence is shaped by the trait-level confidence of the individuals comprising the dyad. We first sought to develop a reliable measure of confidence, which took the form of two matched versions of a difficult 40-item true/false trivia test in which participants answer each question and provide item-level confidence ratings. Having two test versions (Version A and Version B) allowed us to measure the confidence of individuals with different questions from those subsequently faced by dyads—therefore, question novelty was consistent across the individual and dyad contexts, allowing us to focus explicitly on the role of collaboration in dyadic confidence. We then prescreened individuals online with one test version and categorized them as either low or high in confidence based on the observed distribution of individual confidence scores and invited a subset of these participants to the laboratory to measure the effect of individual confidence on dyadic confidence for three types of dyads: Low (composed of two low-confidence individuals), High (composed of two high-confidence individuals), or Mixed (one of each).

This experimental design allows us to explore the following hypothesis and related research questions, motivated by the literature reviewed earlier:

> Hypothesis 1 (H1): The confidence expressed by dyads working together on a difficult true/false trivia test will generally exceed the confidence expressed by the dyad's individual members when working alone.

However, recall that a primary focus of the present work is to explore how individual differences in confidence shape the confidence expressed by dyads. Thus, we ask:

> Research Question 1 (RQ1): Do the confidence gains observed in the dyadic context vary as a function of the trait-level confidence of the individuals who comprise the dyad?

Finally, the present work also seeks to explore the behavioral processes that may underlie any observed effect of individuals' trait-level confidence on dyadic confidence. For this purpose, we used audio and visual recording equipment to capture the dyadic interactions and later coded the recordings for the extent to which each dyad appeared to jointly consider question-relevant knowledge and experiences prior to arriving at a judgment (i.e., engaging in a form of collaborative analysis), as opposed to simply reporting a judgment without appearing to do so (see succeeding discussions for more detail). We also investigate the amount of time that dyads spend arriving at their judgments, which has been shown to predict confidence judgments in prior research (e.g., Pleskac & Busemeyer, 2010; Zakay & Tuvia, 1998). Thus, the present research also asks the following:

> Research Question 2 (RQ2): What characterizes the interaction patterns of dyads that express greater confidence than the confidence of their individual members?

METHOD

In order to test the role of individuals' trait-level confidence in the confidence expressed by dyads, we first developed two matched versions of a general-knowledge trivia test, so that one version could be used to assess individual-level confidence in an earlier online prescreening whereas the other could be used to measure dyadic confidence in the laboratory. Following the development of both test versions, we sought to establish the test reliability of the alternate forms by administering both versions to the same sample of online participants approximately 10 days apart.

Later, we describe the development of the confidence measure in detail before turning to the main study, which involved using the confidence measure to prescreen a large sample of online participants ($N = 686$) for possible inclusion a subsequent in-lab dyadic interaction study.

Development of the confidence measure (online)

Participants

We recruited a total of 308 individuals to participate in one of the three stages during the development of the confidence measure. In *Stage 1*, $n = 100$ participants (70 women and 30 men; mean age (years) = 29.6, $SD = 10.2$) responded to an ad on Craigslist.com offering a $10 gift certificate from Amazon.com for participating in a 45-minute online "trivia quiz." In *Stage 2*, $n = 170$ (120 women and 50 men; mean age = 28.9, $SD = 9.4$) responded to a similar ad offering a $5 gift certificate to Amazon.com for participating in a shorter, 15-minute trivia quiz. In Stage 3, $n = 38$ undergraduates (20 women and 18 men; mean age = 19.5, $SD = 1.4$) responded to flyers posted around campus offering psychology course credit for participating in a 15-minute trivia quiz. In all three stages, participation was limited to high school graduates and native English speakers living in the greater Boston area and willing to be contacted for participation in future studies.

[4]We note that ours is not the only study to instruct dyads to collaboratively reach a confidence estimate (e.g., Allwood & Björhag, 1990).

DOI: 10.1002/bdm

Materials

Tests took the form of general-knowledge true/false tests, administered online, in which individuals selected a response to a statement (i.e., True or False) and then rated their confidence in that response on a 50% to 100% scale, increasing from 50% in increments of 5%. As in previous work, participants were reminded that their confidence could not be less than 50%, because in that case they ought to choose the alternative response (Allwood & Björhag, 1990). All data were collected via the web-based survey service Surveymonkey.com.

Members of our research team initially constructed approximately 160 true/false questions using information from various online sources, including Encyclopedia Britannica Online (britannica.com) and the Guinness Book of World Records (guinnessworldrecords.com). Questions were designed to draw on a diverse set of knowledge, spanning topics such as entertainment, geography, and history (e.g., *Vienna was once the seat of the Holy Roman Empire* [True]; *The Tropic of Capricorn is in the Northern Hemisphere* [False]). Questions were also designed to engender low accuracy, because in light of marked overconfidence on general-knowledge tests, we reasoned that easy questions would engender average confidence estimates that would be too high to be able to meaningfully separate "low confidence" from "high confidence" individuals.

Of the original 160 questions, 120 were chosen and pretested with online participants in Stage 1 to assess accuracy, confidence, and overconfidence engendered ($n = 100$). Ambiguous questions were discarded, and 80 of those remaining were then used to construct four matched 20-item tests. In Stage 2, 170 participants then completed one of these four versions. Using data from the first two phases, we constructed the final two test versions (A and B) that were matched for accuracy, confidence, overconfidence, and content domain. Appendix 1 contains the 80 items comprising the final versions of the trivia test (40 items each).

Procedure

After reaching the survey's web page on Surveymonkey.com, participants first provided their name and demographic data before completing the trivia test. In all cases, question order was fixed, and items were presented in a vertical list on a single, scrollable web page, composed of three columns: one for the question, one for the response (with a drop-down menu to select either "True" or "False"), and one for the confidence rating (with a drop-down menu containing 11 choices, 50% to 100% in increasing increments of 5%). Stage 1 participants completed 120 true/false questions, Stage 2 participants completed 20 true/false questions and 20 of a different type that are not discussed further, and Stage 3 participants completed both versions (A and B) of the final 40-item tests to assess the test–retest reliability of the alternate forms (between-test interval, in days, $M = 9.92$, $SD = 6.91$).

Table 1 shows the characteristics of each version when they were administered to the different samples and to the same sample. Participants' mean confidence across the 40 items was calculated as well as their overall accuracy (i.e., the proportion of correct items). Despite our efforts to match the tests for accuracy, Version A proved significantly more difficult than Version B. However, note that the tests nevertheless generated equivalent confidence levels.

Table 1. Means values (with *SD*s in parentheses) for each test version when administered to different samples (top) (n = 339 for Version A, n = 347 for Version B) and the same sample (bottom) (n = 38 for each)

Sample type	Version A	Version B
Different samples $N = 686$		
Confidence	.69 (.08)	.70 (.08)
Accuracy*	.52 (.08)	.56 (.09)
Overconfidence*	.18 (.11)	.14 (.12)
Same sample, $N = 38$		
Confidence	.66 (.06)	.64 (.06)
Accuracy	.52 (.07)	.54 (.09)
Overconfidence	.14 (.09)	.11 (.09)

*$p < .001$ for mean comparisons.

Test–retest reliability (for Versions A and B)

The scatterplot in Figure 1 shows the relationship between mean confidence on Version A and mean confidence on Version B in our sample of 38 participants. Mean confidence on Version A and mean confidence on Version B were highly correlated at $r = .84$ ($p < .001$). The test–retest correlations for accuracy ($r = -.19$) and overconfidence ($r = .21$) did not indicate significant reliability, nor was confidence significantly predictive of accuracy ($r = .12$). Thus, these patterns suggest that our general-knowledge tests are able to reliably measure individual differences in confidence, in line with previous research reporting correlations for confidence that exceed those for accuracy or overconfidence (Blais et al., 2005; Bornstein & Zickafoose, 1999).

Dyadic confidence on a face-to-face collaborative decision task (in the lab)

Participants

Having developed matched versions of the confidence measure and established their reliability, we recruited a total of

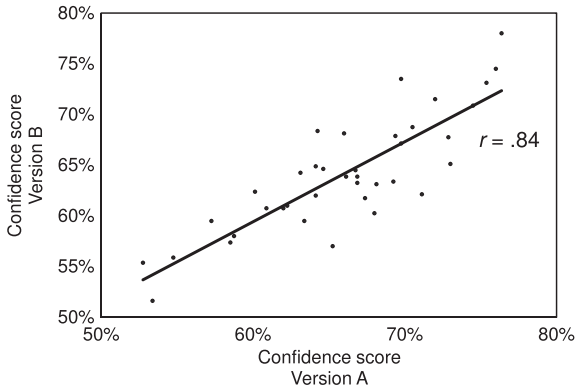

Figure 1. Scatterplot displaying the association between confidence scores (mean of questions) on Versions A and B in the test–retest sample of 38 online participants

DOI: 10.1002/bdm

$N = 686$ to complete one of the two test versions, allowing us to categorize individuals' trait-level confidence as either low or high based on the relative location of their average confidence estimate in the distribution. Individuals with mean confidence scores of 64% or lower were considered "low" and individuals with scores of 75% or higher were considered "high," with cutoffs that corresponded to the terciles observed for the distribution of confidence scores.[5] Overall, 294 men and 392 women comprised the combined sample, ranging in age from 18 to 63 years ($M = 27.68$, $SD = 9.22$). Participation was again limited to high school graduates and native English speakers living in the greater Boston area and willing to be contacted for future studies.

Of the 686 prescreened individuals, a subset of 72 participants (48 women and 24 men; mean age = 30.0, $SD = 10.0$) was recruited via e-mail to complete the alternate form of the confidence measure as part of a dyad working collaborative in the laboratory in exchange for $20 cash.[6] The subset comprised 36 same-sex dyads, 12 in each of the following three conditions: Low, Mixed, and High (described earlier). In each condition, eight dyads were female and four were male (see Table 2 for demographic characteristics of the dyads across type). Dyad types did not differ in average age, years of education, or general cognitive ability (score on a short form of Raven's Advanced Progressive Matrices (RAPM), described subsequently).

There were two additional goals for the laboratory study. First, all dyadic sessions were recorded using digital audio and video equipment, which allowed us to analyze interaction patterns that may help explain the dyads' confidence expressions and accuracy scores, in line with RQ2. The computer program used in the study (see succeeding discussions) also captured the time it took for dyads to arrive at an answer (true or false) and a confidence judgment, which allowed us to explore whether and how decisional duration related to dyadic confidence and accuracy (Pleskac & Busemeyer, 2010; Zakay & Tuvia, 1998). Second, we sought determine the relationship between confidence on our test and general cognitive ability, by correlating lab participants' individual confidence scores with their performance on a short form of the RAPM test (Bors & Stokes, 1998), to further distinguish confidence from potentially related cognitive measures.

Table 2. Select demographic characteristics by dyad type (M (SD))

	Age (years)	Education (years)	RAPM score (out of 12)
Low	29.4 (4.4)	16.9 (1.4)	7.8 (1.4)
Mixed	28.7 (6.5)	17.2 (1.1)	8.0 (1.7)
High	32.0 (11.3)	16.1 (2.1)	8.2 (2.0)

RAPM, Raven's Advanced Progressive Matrices (short form of Bors & Stokes, 1998).
$ps > .10$ for all mean comparisons between dyad types.

Procedure

Upon arriving at the lab, participants first completed a consent form that informed them about the nature of the study and of the use of audio and visual recording equipment; they were then led to a testing room containing this equipment, along with a computer and two chairs, and instructed to sit in a randomly assigned position (left or right). With their partner, they read through the instructions on the computer screen and listened to the experimenter's verbal summary of three key instructions: to alternate control of the mouse for each question; to make every answer and confidence decision jointly, with at least some input from each participant; and to bear in mind the 40-minute time limit for the task.[7]

Participants then worked together to complete the task on an eMac computer (Apple Computer, Cupertino, CA) running PSYSCOPE version 1.2.5 (Cohen, MacWhinney, Flatt, & Provost, 1993) under Classic mode in Mac OS X. Questions were presented in a fixed order identical to that of the online version. Each question trial featured a plain white screen and a numbered statement appearing in black text, beneath which were two gray boxes labeled "True" or "False." Groups were instructed to indicate their chosen alternative by clicking the box corresponding to their answer. Immediately after the response, 11 boxes labeled from 50% to 100% in increasing increments of 5% appeared beneath the text "How confident are you that this answer is correct?" Groups indicated their confidence level by clicking the corresponding box.

After completing the task, participants were led to separate rooms to complete a paper questionnaire and a debriefing. Importantly, this questionnaire included the 12-item short form of the RAPM, which we used to assess general cognitive ability. Each RAPM item presents a 3×3 matrix of visual symbols, with the bottom-right entry missing. The participant chooses from an array of eight options the one that best fits in the empty space. There is only one correct answer for each item; chance performance is 1.5 correct. Participants complete two easy practice items, with feedback, and then receive 15 minutes to complete the test. In total, the laboratory session typically lasted just under 1 hour.

[5]Of the 686 total participants, 609 responded to an ad on Craigslist.com offering a $5 gift certificate to Amazon.com in exchange for participating in a 15-minute "trivia quiz," whereas the remaining 77 individuals responded to an ad placed on Craigslist.com offering a gift certificate to Amazon.com ranging in value between $8 and $28 ($M = \$21.88$, $SD = \$4.47$), depending on choices made in a separate, unrelated task. This longer study was a multiple-components online screening, taking between 45 and 60 minutes, used to select individuals for another group decision-making study that promised to pay at least $10 per hour. In addition to the trivia test, this study included an intertemporal choice (delay discounting) task, a working memory (N-back) task, and a measure of empathy, the results of which are not reported here. The samples did not differ appreciably in age or sex ratio.

[6]All participants with confidence scores in the first or third tercile from the online prescreening were invited by email to participate in the follow-up lab study. Reminder emails targeting group segments with lower response rates (e.g., men with lower confidence scores) were sent periodically until the lab study achieved equal representation of sex and dyad type.

[7]We used a time limit to encourage participants to stay focused on the decision task and chose 40 minutes after pre-testing suggested that dyads would easily be able to finish the task within that timeframe. All dyads completed the task within this timeframe.

DOI: 10.1002/bdm

Behavioral interaction style

We reviewed the videotapes of dyadic interaction and coded them for their general approach to making answer and confidence decisions. Dyads sometimes approached items by first expressing question-relevant knowledge and experiences, arriving at the answer *after* the consideration of evidence that could conceivably justify or lend credence to their choice. Conversely, dyads sometimes settled on an answer *without* verbally expressing any question-relevant knowledge or experience. We call the first, more analytical approach the "Analyze" approach and the second, more action-focused approach the "Act" approach, styles that share some common characteristics with established dual-process models in behavioral decision research, such as the maximize/satisfice distinction (e.g., Diab, Gillespie, & Highhouse, 2008; Schwartz et al., 2002), and System 1 versus System 2 processing (e.g., Sloman, 1996; Stanovich & West, 2000).

Two condition-blind and independent coders categorized every item completed by dyads as predominately exhibiting the Analyze approach, the Act approach, or as not clearly one or the other, using the criteria described earlier. Then, we reduced this trinary coding to binary by combining the intermediate and Act categories into a single category. Thus, questions for which dyads clearly appeared to consider decision-related knowledge were coded Analyze; otherwise, they were coded Act. The coders showed strong inter-rater reliability ($r = .88$), and we analyze the average of their proportion scores in the Results section.

RESULTS

Before turning to the main results of the study, as one measure of confidence realism, we first examined the correlation between confidence and accuracy in our sample of 686 online participants who completed the prescreening procedure. As shown in Figure 2, no significant relationship between confidence and accuracy emerged ($r = -.03$), thus echoing our earlier observation with a much larger sample. We also subtracted participants' mean-level accuracy from their mean-level confidence across items to compute the more common measure of confidence realism, that is, overconfidence/underconfidence. Echoing past observations, on average, individuals were unjustifiably confident given their accuracy (mean overconfidence = 16.3%, $SD = 11.3\%$). We also examined the relationship between confidence and general cognitive ability, as measured by the 12-item Raven's test, in the 72 participants who participated in the lab study. We found that general cognitive ability did not significantly predict confidence ($r = .12$).[8]

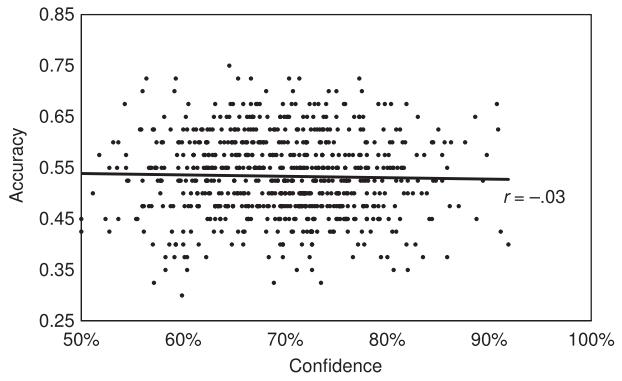

Figure 2. Scatterplot displaying the association between confidence (mean of questions) and test accuracy (i.e., proportion of correct items) in our sample of 686 online participants

Individual versus dyadic confidence

Recall that we expected that, overall, the confidence expressed by dyads working collaboratively would exceed that expressed by individuals when they worked on a matched version of the general-knowledge test alone (H1). As expected, dyadic confidence exceeded the average of individuals' confidence: 74.1% versus 69.7%, $t(35) = 4.02$, $p < .001$. We also computed bias scores to test whether groups were more or less overconfident than individuals. Although groups also demonstrated more overconfidence than did individuals, this difference was not significant: 18.7% versus 15.6%, $t(35) = 1.49$, $p = .13$.

Individual versus dyadic confidence by dyad type

Recall that we were primarily interested in testing whether the gains exhibited by participants in the dyadic versus individual setting varied as a function of dyad type, that is, whether dyads were composed of two low-confidence individuals (Low), two high-confidence individuals (High), or one of each (Mixed; RQ1). Results revealed a different pattern across dyad types, with the greatest confidence gains exhibited by dyads composed of two low-confidence individuals (Figure 3). Specifically, Low dyads expressed significantly greater confidence in the dyadic setting compared with when working alone: 70.8% versus 60.4%, $t(11) = 8.55$, $p < .001$. Mixed dyads also expressed significantly greater confidence in the dyadic setting but showed a smaller gain than did the Low dyads: 73.3% versus 69.9%, $t(11) = 2.73$, $p = .02$. High dyads, in contrast, showed no significant difference in confidence between the individual and dyadic setting: 78.2% as dyads versus 78.8% as individuals.

Also apparent in Figure 3 is that dyadic confidence for High dyads (78.2%) was significantly greater than the confidence expressed by Mixed (73.3%) and Low dyads (70.8%; ps $< .05$ and $< .01$, respectively); however, Mixed and Low dyads did not differ significantly ($p > .10$). This pattern suggests that on the present task, group confidence expressions are not the result of a simple averaging of individual judgments nor do they gravitate toward the judgment of the most confident individual in the group, as recent work has found

[8]We later replicated this lack of a positive relationship between trait confidence and general cognitive ability in a separate online study using the 40-item confidence measures developed here and a different measure of general cognitive ability: the MiniCog Rapid Assessment Battery (Shephard & Kosslyn, 2005), a set of nine brief cognitive tests designed to cover the domains of working memory, attention, and problem-solving. Mean performance (measured as percentage of correct responses) across all nine MiniCog Rapid Assessment Battery tasks correlated $r = .01$ with confidence on version A ($n = 111$) and $r = -.29$ with confidence on version B ($n = 120$).

DOI: 10.1002/bdm

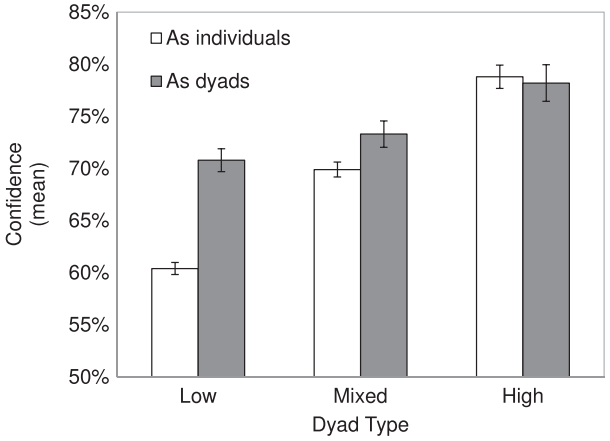

Figure 3. Confidence (mean) as individuals and as dyads, by dyad type. Error bars represent mean standard errors

(Koriat, 2012). Instead, it appears that dyads composed of a high-confidence member and a low-confidence member generate judgments that are more similar to those generated by two low-confidence individuals than to those generated by two high-confidence individuals.

Accuracy

We also analyzed the accuracy of participants' decisions, by dyad type and across the individual and group settings (Table 3). Results revealed no accuracy differences across dyad types (Low = 55%, Mixed = 56%, High = 55%) or individual and dyadic contexts (individual = 54% correct, group = 55%). Finally, confidence and accuracy were not significantly correlated at the dyad level ($r = .24$, $p > .15$). Within dyad types, the confidence-accuracy correlations were $r = -.05$, .21, and .52 for the Low, Mixed, and High dyad types, respectively; owing in part to the low group *N*s, none of these correlations were significant at the .05 level (ps $> .08$).

Response times

We also analyzed the response time (RT) for each answer decision and confidence judgment made by every dyad (recorded by computer software), as well their social interaction for the Analyze versus Act style described earlier. Table 3 shows mean RT for the answer decision and the confidence judgment, by condition. As others have observed (Allwood & Björhag, 1990), dyads spent little time discussing the confidence judgment relative to the answer decision, despite being specifically instructed to confer on both the answer and confidence decisions. Interestingly, Mixed dyads took significantly longer than both Low and High dyads to settle on an answer (in seconds): Mixed = 35.2, Low = 23.1, High = 25.9 (Mixed versus Low: $t(22) = 2.47$, $p = .02$; Mixed versus High: $t(22) = 2.05$, $p = .05$). Moreover, reiterating longstanding observations in the literature (e.g., Henmon, 1911; Johnson, 1939; Zakay & Tuvia, 1998), item-level analysis revealed that longer times for groups to decide on answers (true/false) were significantly associated with lower expressions of confidence, a pattern that emerged for all dyad types (Low: $\rho(461) = -.11$; Mixed: $\rho(459) = -.10$; High: $\rho(463) = -.11$; all ps $< .05$; in contrast, answer response times were not significantly associated with accuracy or overconfidence at the item level, all $|\rho|$s $< .08$, *ns*).

Behavioral interaction style

Evaluation of analyze versus act style revealed a difference across conditions in the extent to which dyads discussed question-relevant knowledge and experiences before arriving at an answer (Analyze), versus settling on an answer *without* discussing any question-relevant knowledge supporting their choice (Act).

Table 3 also displays key results from the Act/Analyze coding analysis. As expected, Analyze score (the proportion of items on which the dyad demonstrated the Analyze approach) was positively associated with group accuracy, $r = .51$, $p < .01$. Analyze score was also positively associated with group confidence, but this correlation was only marginally significant, $r = .31$, $p = .08$. Thus, the more a dyad approached a given item by first discussing judgment-relevant knowledge or experiences, appearing to settle on the final answer *through* the consideration of that information, the better that dyad performed (and, to a lesser extent, the higher its confidence). Analyze score also differed by dyad type, with Low = .32, Mixed = .48, and High = .37; Mixed dyads demonstrated the Analyze approach significantly more often than Low dyads ($p < .05$), but the comparison with High dyads was only marginally significant ($p = .08$).

DISCUSSION

Groups that perform analytic tasks are not called upon only to draw conclusions, make projections, and plan courses of

Table 3. Summary of accuracy, confidence, and behavioral findings by dyad type (*SD*s in parentheses)

Dyad type	Accuracy (%)	Confidence (%)	Answer decision time (*M*) (seconds)	Confidence decision time (*M*) (seconds)	Analyze (%)	Act (%)
Low	55.0 (9.0)	70.8 (3.8)	23.1 (10.8)	5.1 (2.8)	31.7 (16.4)	68.3 (16.4)
Mixed	56.1 (5.9)	73.3 (4.4)	34.3 (11.3)	6.3 (2.8)	47.8 (12.3)	52.2 (12.3)
High	55.1 (10.4)	78.2 (6.1)	25.9 (8.5)	5.2 (2.3)	37.3 (13.4)	62.7 (13.4)
*N*s	36	36	36	36	33	33

Analyze and Act columns depict the percentage of trials on which dyads arrived at an answer having first discussed question-relevant knowledge or *without* having done so, respectively (as determined by two independent coders). Mixed dyads showed the Analyze approach significantly more often than did Low dyads ($p = .02$) and marginally more often than did High dyads ($p = .08$; note: Three dyads—two Mixed and one High—were excluded owing to technical difficulties with recording equipment, leaving $n = 33$ for the Act/Analyze analysis).

DOI: 10.1002/bdm

action; often, they must also predict and express the likelihood that their estimates are correct and that their plans will succeed. In this study, we found that dyad-level confidence on a difficult task exceeded individual confidence while bearing no relationship with accuracy. The magnification of confidence in dyads may pose a special danger because the consumers of confidence judgments may draw the unwarranted inference that a more confident group is likely to be a more accurate group. Of course, higher confidence is justified by commensurate gains in accuracy, but unlike studies finding greater accuracy in groups than individuals (Allwood & Granhag, 1996; Allwood, Granhag, & Johansson, 2003; Littlepage, 1991; Sniezek & Henry, 1989), our participants were no more accurate when they completed the task in dyads rather than individually, which might be explained by the intentionally difficult nature of the questions we employed. At the same time, it is reasonable to expect that real analytic groups sometimes confront equally challenging forced-choice scenarios in which they must collectively reach decisions and generate uncertainty estimates, in situations where the "correct" course of action is exceedingly difficult to determine (e.g., whether or not to attack a suspected terrorist hideout; see further discussion later). Our results suggest that in these cases, confidence magnification among low-confidence individuals working together might be especially likely.

Perhaps most importantly, we observed that not all dyads gain in confidence equally. While overall dyadic confidence was significantly greater than individual confidence, the majority of this effect was attributable to the dramatic gains in confidence demonstrated by two low-confidence individuals working together. Mixed dyads, with one low-confidence individual and one high-confidence individual, also increased significantly in confidence, whereas dyads composed of two high-confidence individuals showed no increase. Put another way, it appears that low-confidence dyads experienced the greatest amount of process loss from working together: Whereas low-confidence participants exhibited merely 6.4% overconfidence as individuals, this figure rose to 15.8% when they worked collaboratively with another low-confidence individual.

One possible explanation for the differential gains in confidence observed across dyad types is with regard to the ability of group settings to reduce individual feelings of uncertainty. Social comparison theory (Festinger, 1954) posits that individuals are motivated to assess the validity of their opinions by comparing them to those held by others, in the absence of other non-social, "physical" means for doing so. Social comparison processes are thus more likely to operate in groups performing tasks that are more judgmental, as compared with intellective, in nature (Laughlin & Earley, 1982), as when individuals in a jury scenario adopt a higher threshold for finding a suspect guilty ("beyond a reasonable doubt") after participating in a group discussion (Magnussen, Eilertsen, Teigen, & Wessel, 2014). Although our trivia tests would seem to qualify as highly intellective, recall that the test items were designed to elicit low-accuracy rates, positioning the task closer to the judgmental end of the intellective–judgmental continuum (i.e., "eureka" solutions were very unlikely). To the extent that the group setting serves this social-informative function, it is reasonable to expect that compared with high-confidence individuals, low-confidence individuals would have more to gain in terms of reducing their feelings of uncertainty, perhaps leading to greater gains in post-discussion (dyadic) confidence.

This pattern of results may also reflect, at least in part, statistical regression to the mean (Kahneman & Tversky, 1977), which is relevant whenever individuals are selected for extreme performance on some measure at Time 1, as they were in the present study. However, we do not believe it can fully account for the pattern of results observed here. First, if the confidence gains of low-confidence dyads were simply due to statistical regression, we would expect high-confidence dyads to show a corresponding fall in confidence at the group level; this did not occur. Second, the test–retest reliability for the alternate forms was $r = .84$, demonstrating that an individual's confidence score on one version was highly predictive of their score on the other version. A ceiling effect for high-confidence dyads also seems unlikely, because in both the individual and group contexts, the average confidence for high-confidence participants was around 78%, leaving about 22% headroom. It remains possible, however, that there could be a socially imposed ceiling whereby individuals tend to avoid displays of high confidence (say, above 80%), perhaps especially on difficult forced-choice judgment tasks like the present one that may involve guessing or the perception of guessing.

Besides our main finding that confidence gains demonstrated by dyads differ as a function of members' trait-level confidence, the present results contribute to the confidence literature in a number of ways. The observation that the confidence expressed by Mixed dyads was indistinguishable from that expressed by Low dyads might at first appear incompatible with recent work suggesting that dyads benefit by conforming to the opinion of the higher confidence individual (Bahrami et al., 2010; Koriat, 2012). However, our study differs from those studies important ways, most notably in that our test was designed to be especially difficult; indeed, our participants performed just slightly better than chance, and confidence was uncorrelated with accuracy.[9] We contend that society is routinely faced with decisions of the kind we study here: "coin flip" decisions—essentially, guesses (see later discussions)—made collaboratively by groups on the basis of a joint expression of subjective confidence, with little or no immediate feedback regarding decision accuracy. From this perspective, our finding that the presence of just one cautious individual in a Mixed dyad helped to mitigate the (unjustified) confidence gains exhibited by homogenous high-confidence dyads may carry optimistic implications for improving collaborative performance through thoughtful group composition, a long-standing focus of research in numerous disciplines, including psychology, management, and communication (e.g., Barry & Stewart, 1997; Fisher & Ellis, 1980; Hackman,

[9]We also instructed dyads to alternate control of the computer mouse, which may have mitigated any tendency for higher confidence individuals to dominant the dyadic interaction.

DOI: 10.1002/bdm

2002). In terms of response times, Mixed dyads also took significantly longer to settle on answer decisions than did either of the two other dyad types, possibly reflecting the added process challenges faced by more heterogeneous dyads. However, we observed no difference in the time spent on confidence decisions across dyad types, perhaps partly because dyads allocated only a small fraction of their time to confidence judgments relative to answer decisions (about 25% as much), and so it appears that the confidence combination process occurs prior to the actual discussion of the confidence rating (Allwood & Björhag, 1990).

In addition, our results help illuminate the interpersonal communication processes that underlie the differences observed across dyad types. Mixed dyads took the Analyze approach to settling on an answer more often than the other dyads; that is, they were more likely to verbally express and consider decision-relevant knowledge or experiences that could justify or lend credence to whichever answer they eventually chose (i.e., true or false). Although this point is speculative, we suggest that the heterogeneity in trait-level confidence may be responsible for the prolonged discussion period exhibited by Mixed dyads, which in turn was associated with lower dyadic confidence expressions in our data. This finding also suggests that even for extremely difficult forced-choice decisions, accuracy may be improved by a process intervention—namely, instructing groups to consider relevant evidence prior to and as a means to settling on a decision. It is important to note, however, that our data relating behavioral interaction style to group outcomes are correlational in nature, limiting our ability to draw causal inferences here.

We also note some of the limitations of this study. We utilized convenience samples of individuals living in the greater Boston area in developing matched versions of our confidence measure and prescreening individuals for the subsequent in-lab dyad task. As a result, the results reported here may not be fully generalizable to the population at large, although the fact that we replicated some previous observed findings (i.e., greater confidence expressed by dyads than individuals, more time spent on answer decisions than confidence decisions in a group setting; Allwood & Björhag, 1990; Allwood & Granhag, 1996) might mitigate this concern. At the same time, however, our findings deviate from some previous work in that dyads were no more accurate than individuals, which we suspect is due to the intentionally difficult nature of the questions we employed. In addition, the present work focused exclusively on the confidence expressions of unacquainted dyads interacting face-to-face. Thus, it is unclear whether similar processes may unfold in computer-mediated communication, which affords different challenges and opportunities than face-to-face interactions (e.g., Rice, 1987; Tidwell & Walther, 2002). Future research may fruitfully explore whether similar processes emerge among computer-mediated dyads, in addition to the role of acquaintance status in these effects, given that collaborative dyads frequently have a shared history of joint decision making that may shape the processes examined here.

Moreover, the specific nature of the decision task employed here poses some additional limitations. For example, a number of these questions referenced events that were unlikely but true, and therefore, unrepresentative of many types of decisions that people face in their everyday lives. Thus, although we have primarily interpreted the confidence expressed by individuals and dyads as reflecting genuine confidence in their factual knowledge, the type of questions used here makes it likely that our participants sometimes had little knowledge to guide their responses and were sometimes engaged in pure guessing. Although such situations may seem somewhat unusual at first glance, we contend that collaborative groups routinely face highly uncertain, consequential decisions in which factual knowledge is limited and where guesswork is bound to play a significant role. For example, Bergen (2012) describes how the decision by high-ranking U.S. personnel to storm the compound in Pakistan where Osama Bin Laden was ultimately killed was rife with uncertainty about whether the terrorist leader was actually there. As part of the decision-making process, officials repeatedly discussed their personal level of confidence that Bin Laden was on the premises.

> The analysts believed this with varying degrees of certainty, with most estimating the probability at 80 percent. The lead analyst, John, was still at about 90 percent, while Michael Morell, the deputy director of the CIA, was at 60 percent (p. 133).

A fresh "red team" of intelligence analysts was brought in to produce independent estimates, which ranged from 40% to 60%. Despite these far-from-certain levels of confidence, the collective decision was taken to go forward. Confidence is not always realistic in such high-stakes cases; for example, the Bush administration's confidence in 2003 that Iraq possessed weapons of mass destruction proved less than prescient.

Although we believe that guessing among our participants helped to simulate real-world decision tasks that involve the aggregation of individual-level confidence judgments, at the same time, we acknowledge that it raises important considerations for the interpretation of the present results. For example, our main finding that low-confidence dyads gained in confidence more than any other dyad type may be partly explained by the prevalence of guessing: If low-confidence individuals agree on a particular answer choice, it may be reasonable for them to increase their confidence judgment, whereas if they disagree, there is little space for their (already low) confidence to decline much further (a floor effect). On the other side, if high-confidence individuals disagree, it may be reasonable for them to *decrease* their confidence judgment, whereas if they share the same guess, there is little space for their (already high) confidence to increase much further (a ceiling effect).[10] More generally, given the cultural script for "tricky" true/false tests that feature a devious test-maker who crafts questions that are plausible but false

[10]We thank an anonymous reviewer for suggesting the possible role of guessing in the observed effects.

DOI: 10.1002/bdm

(e.g., *The Eiffel Tower was once scheduled to be demolished in 1919*; the correct year is 1909) as well as implausible but true (e.g., *The longest fit of continuous hiccupping lasted more than 20 years*), confidence expressions might also encompass participants' subjective feelings about their ability to detect such trickery, as well as the norm that roughly half of such questions are true and half are false (as was indeed the case here). Given that our participants may have been engaging in these kinds of meta-cognitive strategies, we caution against generalizing too far beyond this particular decision context.

Overall, this work highlights the interactive nature of individual traits in group decision making, with the findings that dyad confidence and decision time varied as a function of the individual confidence levels of dyad members. Individual differences in confidence have long been overlooked in investigations of group confidence, and these findings suggest that knowing the trait confidence of individuals would help those charged with selecting group members. This would require a validated instrument for measuring individual confidence. While we were able to develop a reliable measure of individual differences in confidence, this was accomplished using a trivia test that is somewhat idiosyncratic with respect to the present time period and cultural background of young adult Americans. This test was well suited for both individuals and dyads, as it allowed for debate over correct answers and confidence levels, and it included a diverse selection of knowledge domains, decreasing the likelihood that one's perceived competency in any single knowledge domain had undue influence on confidence judgments. However, future research should attempt to develop psychometrically reliable and valid measures of confidence that are less culture-bound. Organizations engaged in making decisions and recommendations with associated expressions of confidence should be aware that the group context may significantly inflate confidence even when the group is facing a novel problem or task, and that the trait confidence levels of group members can influence the size of this inflation effect.

APPENDIX

The following are the two matched general-knowledge trivia tests that were constructed by researchers for use in this study. Correct responses (T = True, F = False) are given as of the time when the studies reported in this article were conducted (2004–2005).

Version A

1. Chicago has a larger population than Toronto. [T]
2. The bicycle was invented in Scotland. [T]
3. Mozart wrote *The Magic Flute* in the year that he died. [T]
4. A 2005 Mercedes-Benz E320 is more expensive than a 2004 Lexus LS 430, when both are equipped with their basic features. [F]
5. More than 700,000 people died in the American Civil War. [F]
6. Paul McCartney once performed for a crowd of over 180,000 people. [T]
7. Macadamia nuts are less than 37% fat. [F]
8. Dogs were domesticated before horses. [T]
9. When counting all parts that are integral to the overall architectural structure, the tallest building in the world is in Taiwan. [T]
10. The cost of living is higher in Los Angeles than it is in San Francisco. [F]
11. Jaundice is the condition which leads to yellow to green discoloration of the skin. [T]
12. There are over 1,000 fountains in the Bellagio Hotel in Las Vegas, NV. [T]
13. Irving Berlin wrote more scores for movies than he did for Broadway shows. [F]
14. Melos and Naxos are both Greek Isles [T].
15. Japan is the nation with the highest life expectancy. [F]
16. The top speed of the first automobile was 8 miles/hour. [F]
17. Dave Thomas, founder of Wendy's, appeared in over 1000 commercials for the restaurant chain. [F]
18. Voting is mandatory in Mexico. [T]
19. Over $180 billion of damage was caused worldwide by natural disasters in 1995. [T]
20. *Star Wars* is ranked 3rd in United States box office earnings, as of July 2004. [F]
21. Combined, all the planets in our solar system contain 2% of the matter found in the whole solar system. [F]
22. China borders exactly 15 nations. [F]
23. Fewer than eight European countries contain part of the Alps. [F]
24. Faulkner's *The Sound and the Fury* was published in the same year as Hemingway's *The Sun Also Rises.* [F]
25. The largest window in the world is in France. [T]
26. The United Kingdom is the second most educated country in the world, based on average years of education. [F]
27. A neutron star has an average diameter of 10 kilometers, but has the same mass as our sun. [T]
28. The Tropic of Capricorn is in the Northern Hemisphere. [F]
29. Vienna was once the seat of the Holy Roman Empire. [T]
30. The Eiffel Tower was once scheduled to be demolished in 1919. [F]
31. Air France carries more passengers per year than British Airways. [F]
32. *Hamlet* is Shakespeare's longest play. [T]
33. The longest fit of continuous hiccupping lasted more than 20 years. [T]
34. The X-ray was discovered in 1915. [F]
35. Emily Dickinson's first book of poems was published posthumously. [T]
36. As of July 2004, the oldest living man and the oldest living woman were both American citizens. [T]
37. A sculpture of Michael Jackson and his pet monkey, Bubbles, was sold for $1,000,000 in 1988. [F]
38. The world's largest shopping mall, based on square feet, is in Canada. [T]

DOI: 10.1002/bdm

39. More Americans die of colon cancer than any other form of cancer. [F]
40. Otto von Bismarck was both the prime minister of Prussia and the chancellor of the German Empire. [T]

Version B

1. San Jose, CA, has a larger population than Dallas, TX. [F]
2. The Home Depot had more revenue in 2003 than Verizon Communications. [F]
3. More rulers of the United Kingdom have been named Charles than George. [F]
4. More than 10% of the world population speaks English as their first language. [F]
5. The oldest age ever reached by an alligator is 66. [T]
6. France has a larger Muslim population than Protestant population. [T]
7. The Statue of Liberty weighs over 70 million pounds. [F]
8. "War communism" was an economic policy specific to China. [F]
9. The first mental hospital was established after the Black Plague. [F]
10. The New England Patriots have played in 4 Super Bowls. [T]
11. Fidel Castro was born in 1931. [F]
12. Prince William has over 72 fan clubs. [F]
13. Washington, D.C. has a larger population than Wyoming. [T]
14. Stevie Wonder achieved his first #1 album at age 16. [F]
15. More murders per thousand people are committed in Colombia than in any other country. [T]
16. More parakeets were kept as pets in America in 2003 than all the pet rabbits, gerbils, hamsters and small rodents combined. [T]
17. Saturn has a higher density than Jupiter. [F]
18. Both solo and duet synchronized swimming were eliminated from the Olympic games in 1996. [T]
19. The state of Maine is larger in area than Austria. [T]
20. The Arctic Ocean is smaller in area than the United States. [F]
21. The electron was the first subatomic particle discovered. [T]
22. Prisoner of war exchange from the Iran–Iraq war was not complete until 2001. [F]
23. The elevator was invented before the escalator. [T]
24. The first Blockbuster Video store opened in 1980. [F]
25. Over 1000 earthquakes and other seismic events take place in Japan every year. [T]
26. Barbara Streisand has won over 25 Oscars, Emmys, Grammys, Golden Globes, and Tonys combined. [F]
27. The United States leads the world in beer consumption per thousand people. [F]
28. Squirrels can reach a top speed of 12 miles per hour. [T]
29. The O.J. Simpson murder trial ended in 1993. [F]
30. Each year more Americans are killed or injured in accidents at home or work than were killed or injured in the Vietnam War. [T]
31. Dell Computers had more revenue in 2003 than Hewlett-Packard, Inc. [F]
32. Asia has a larger urban population than rural population. [F]
33. The Great Wall of China is visible from outer space. [T]
34. Jerry Seinfeld was paid $1 million per episode in the final season of *Seinfeld*. [T]
35. Only one of the ancient wonders of the world is known to still exist. [T]
36. The greatest snowfall ever from a single snowstorm occurred at Mt. Rainier, in Washington State. [F]
37. France is the country most visited by tourists. [T]
38. The first spacewalk by a woman was performed in 1984. [T]
39. The first beauty contest in the United States was held in 1855. [T]
40. The largest forest in the world is found in Canada. [F]

ACKNOWLEDGEMENTS

The authors thank Meg Gerbasi, Thomas Jerde, Benjamin Bibler, and Sean Bennett for their help in preparing study materials, and Eugene Burnstein for his valuable comments on an earlier draft. A special thanks to Stephen Kosslyn, whose collaboration with Richard Hackman inspired this work. This research was supported by a research grant from the National Science Foundation (NSF) (REC-0106070, CFDA no. 47.076) and by Fred Ambrose and the Intelligence Technology Innovation Center at the Central Intelligence Agency.

REFERENCES

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review*, *68*, 33–45.

Allwood, C. M., & Björhag, C. G. (1990). Are two judges better than one? On the realism in confidence judgments by pairs and individuals. In Caverni, J. P., Fabre, J. M., & Gonzalez, M. (Eds.), *Cognitive biases* (pp. 443–463). Amsterdam: Elsevier.

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

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

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

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception and Psychophysics*, *55*, 412–428.

Barry, B., & Stewart, G. L. (1997). Composition, process, and performance in self-managed groups: The role of personality. *Journal of Applied Psychology*, *82*, 62–78.

Bergen, P. L. (2012). *Manhunt: The ten-year search for Bin Laden from 9/11 to Abbottabad*. New York: Crown.

Blais, A. R., Thompson, M. M., & Baranski, J. V. (2005). Individual differences in decision processing and confidence judgments

DOI: 10.1002/bdm

in comparative judgment tasks: The role of cognitive styles. *Personality and Individual Differences*, *38*, 1707–1713.
Bornstein, B. H., & Zickafoose, D. J. (1999). "I know I know it, I know I saw it": The stability of the confidence-accuracy relationship across domains. *Journal of Experimental Psychology: Applied*, *5*, 76–88.
Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, *58*, 382–398.
Buratti, S., Allwood, C. M., & Johansson, M. (2014). Stability in the metamemory realism of eyewitness confidence judgments. *Cognitive Processing*, *15*, 39–53.
Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Mac computers. *Behavior Research Methods, Instruments and Computers*, *25*, 257–271.
Diab, D. L., Gillespie, M. A., & Highhouse, S. (2008). Are maximizers really unhappy? The measurement of maximizing tendency. *Judgment and Decision Making*, *3*, 364–370.
Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology*, *58*, 568–581.
Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117–140.
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.
Fisher, B. A., & Ellis, D. G. (1980). *Small group decision making: Communication and the group process*. New York: McGraw-Hill.
Fleischer, A. (2003, April 10). Transcript of White House press briefing. Retrieved 5 April 2007 from: http://www.whitehouse.gov/news/releases/2003/04/20030410-6.html.
Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.
Hackman, J. R. (2002). *Leading teams: Setting the stage for great performances*. Cambridge, MA: Harvard Business School Press.
Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making*, *5*, 467–476.
Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*, 186–201.
Janis, I. (1972). *Victims of groupthink*. Boston: Houghton-Mifflin.
Johnson, D. M. (1939). Confidence and speed in the two-category judgment. *Archives of Psychology*, *241*, 1–52.
Jonsson, A. C., & Allwood, C. M. (2003). Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality and Individual Differences*, *34*, 559–574.
Kahneman, D., & Tversky, A. (1977). *Intuitive prediction: Biases and corrective procedures*. McLean, VA: Decisions and Designs Inc.
Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–655.
Kleitman, S., & Stankov, L. (2001). Ecological and person-driven aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology*, *15*, 321–341.
Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*, 161–173.
Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 107–118.
Koriat, A. (2012). When are two heads better than one and why? *Science*, *336*, 360–362.
Laughlin, P. R., & Earley, P. C. (1982). Social combination models, persuasive arguments theory, social comparison theory, and choice shift. *Journal of Personality and Social Psychology*, *42*, 273–280.
Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*, 159–183.
Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, *26*, 149–171.
Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In Kahneman, D., Slovic, P., & Tversky, A. (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge: Cambridge University Press.
Littlepage, G. E. (1991). Effects of group size and task characteristics on group performance: A test of Steiner's model. *Personality and Social Psychology Bulletin*, *17*, 449–456.
Magnussen, S., Eilertsen, D. E., Teigen, K. H., & Wessel, E. (2014). The probability of guilt in criminal cases: Are people aware of being 'beyond reasonable doubt'? *Applied Cognitive Psychology*, *28*, 196–203.
Myers, D. G., & Lamm, H. (1975). The polarizing effect of group discussion: The discovery that discussion tends to enhance the average prediscussion tendency has stimulated new insights about the nature of group influence. *American Scientist*, *63*, 297–303.
Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., et al. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology*, *129*, 257–299.
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*, 864–901.
Rice, R. E. (1987). Computer-mediated communication and organizational innovation. *Journal of Communication*, *37*, 65–94.
Schraw, G. (1997). The effect of generalized metacognitive knowledge on test performance and confidence judgments. *Journal of Experimental Education*, *65*, 135–146.
Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, *83*, 1178–1197.
Shephard, J. M., & Kosslyn, S. M. (2005). The MiniCog Rapid Assessment Battery: Developing a "blood pressure cuff for the mind.". *Aviation, Space and Environmental Medicine*, *76*, B192–B197.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.
Sniezek, J. A. (1992). Groups under uncertainty: An examination of confidence in group decision making. *Organizational Behavior and Human Decision Processes*, *52*, 124–155.
Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*, 1–28.
Stankov, L. (1998). Calibration curves, scatterplots and the distinction between general knowledge and perceptual tasks. *Learning and Individual Differences*, *10*, 29–50.
Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, *21*, 971–986.
Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, *23*, 701–717.
Stoner, J.A.F. (1961). A comparison of individual and group decisions involving risk. Unpublished Master's Thesis, Massachusetts Institute of Technology.
Tidwell, L. C., & Walther, J. B. (2002). Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human Communication Research*, *28*, 317–348.
Turner, M. E., & Pratkanis, A. R. (1998). Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. *Organizational Behavior and Human Decision Processes*, *73*, 105–115.

DOI: 10.1002/bdm

West, R. F., & Stanovich, K. E. (1997). The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin and Review*, *4*, 387–392.

Woodward, B. (2004). *Plan of attack*. New York: Simon & Schuster.

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

Zakay, D., & Tuvia, R. (1998). Choice latency times as determinants of post-decisional confidence. *Acta Psychologica*, *98*, 103–115.

Authors' biographies:

Jonathon P. Schuldt is an Assistant Professor of Communication at Cornell University. He received his doctorate in social psychology from the University of Michigan. His research examines the communication processes that influence everyday judgment and decision making.

Christopher F. Chabris is an Associate Professor of Psychology and Co-Director of the Neuroscience Program at Union College. He received his doctorate in psychology from Harvard University. His research examines various topics related to the role of individual differences in cognition and decision making.

Anita Williams Woolley is an Associate Professor of Organizational Behavior and Theory at the Tepper School of Business, Carnegie Mellon University, where she has been a faculty member since 2008. She received her doctorate in organizational behavior from Harvard University in 2003 and then worked as a postdoctoral associate on the Group Brain Project with J. Richard Hackman and Stephen Kosslyn until 2008. Her research interests center on team collaboration, performance, and collective intelligence.

J. Richard Hackman was the Edgar Pierce Professor of Social and Organizational Psychology at Harvard University until his death in 2013. He received his bachelor's degree in mathematics from MacMurray College and his doctorate in social psychology from the University of Illinois. He taught at Yale for 20 years before moving to Harvard in 1986. He conducted research on a variety of topics in social and organizational psychology, including team performance, leadership effectiveness, and the design of self-managing teams and organizations.

Authors' addresses:

Jonathon P. Schuldt, Department of Communication, Cornell University, Ithaca, NY, USA.

Christopher F. Chabris, Department of Psychology, Union College, Schenectady, NY, USA.

Anita Williams Woolley, Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA.

J. Richard Hackman, Department of Psychology, Harvard University, Cambridge, MA, USA.

DOI: 10.1002/bdm