Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Gender Composition and Group Confidence Judgment: The Perils of All-Male Groups

Steffen Keck, Wenjie Tang

To cite this article:
Steffen Keck, Wenjie Tang (2017) Gender Composition and Group Confidence Judgment: The Perils of All-Male Groups. Management Science

Published online in Articles in Advance 20 Dec 2017

. https://doi.org/10.1287/mnsc.2017.2881

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–22
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Gender Composition and Group Confidence Judgment: The Perils of All-Male Groups

Steffen Keck,[a] **Wenjie Tang**[b]

[a] Department of Business Administration, University of Vienna, 1090 Vienna, Austria; [b] Business School and Institute of Operations Research and Analytics, National University of Singapore, Singapore 119245

Contact: steffen.keck@univie.ac.at, http://orcid.org/0000-0003-0090-3456 (SK); wenjie.tang@nus.edu.sg, http://orcid.org/0000-0003-3322-4804 (WT)

Received: November 23, 2015
Revised: January 16, 2017
Accepted: June 9, 2017
Published Online in Articles in Advance: December 20, 2017

https://doi.org/10.1287/mnsc.2017.2881

Abstract. We explore the joint effects of group decision making and group gender composition on the calibration of confidence judgments. Participants in two laboratory experiments, individually and in groups of three, stated confidence interval estimates for general-knowledge questions and for financial forecasts. Across both studies, our results reveal that groups with at least one female member are significantly better calibrated than all-male groups. This effect is mediated by the extent to which group members share opinions and information during the group discussion. Moreover, we find that compared to a statistical aggregation of individual confidence intervals, group discussions have a neutral or positive effect on the quality of confidence judgments for groups with at least one female group member; in contrast, group discussion actually harms confidence calibration for all-male groups. Overall, our findings indicate that compared to all-male groups, even the inclusion of a small proportion of female members can have a strong effect on the quality of group confidence judgment.

History: Accepted by Yuval Rottenstreich, judgment and decision making.
Funding: This work was supported by the Foundation at IE Business School and the University of Vienna.
Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mnsc.2017.2881.

Keywords: overconfidence • group judgments • gender diversity • group deliberation

1. Introduction

Decision makers in organizations frequently need to cope with severe uncertainty. In situations like these, adequate levels of confidence may be just as important for organizational performance as the decisions' actual quality (e.g., Sniezek 1992, Sniezek and Henry 1989). However, a large number of studies have shown that individuals' level of confidence is in fact not well calibrated. Instead, most people are systematically overconfident; that is, they hold an excessive certainty concerning the correctness and precision of their beliefs, judgments, and forecasts (e.g., Lichtenstein and Fischhoff 1977, Russo and Schoemaker 1992, Soll and Klayman 2004). Such miscalibration has been shown to have an important effect on decision making in organizations. For example, overconfident investors take too many risks, earn lower average returns, and underdiversify their portfolios (e.g., Barber and Odean 2002, Biais et al. 2005). Similarly, results by Ben-David et al. (2013) show that firms with overconfident chief financial officers pursue more aggressive corporate policies such as larger investments and higher debt levels, which exposes their companies to excessive risk. Confidence judgments are also widely used in decision analysis; therefore, overconfidence in managers' judgments can have serious consequences for the quality of decisions that are based on this methodology (Clemen 1996).

In the last several decades, following an ongoing shift from organizing work around individual jobs to team-based work structures, a large number of important judgments and decisions in organizations are now made by groups rather than by individuals (e.g., Kozlowski and Bell 2003). Furthermore, due to the increasing diversity of the workforce as a whole, teams in organizations are not only becoming more important, but also have become more diverse in terms of demographic categories such as gender, age, and ethnicity (e.g., Triandis et al. 1994). In particular, even though all-male groups are still a ubiquitous phenomenon in many areas such as upper management, boards of directors, the financial sector, and certain areas of academia, mixed-gender groups have become more and more common in the workplace (e.g., Heilman 2012).

Even though a substantial number of studies have explored the effects of gender diversity on group performance in a variety of settings, such as small work teams (e.g., Jehn et al. 1999, Wegge et al. 2008), top management teams (e.g., Dezsö and Ross 2012, Krishnan

and Park 2005), boards of directors (e.g., Adams and Ferreira 2009, Post and Byron 2015), and student competitions (e.g., Apesteguia et al. 2012, Hoogendoorn et al. 2013), the potential effect of gender diversity on a group's susceptibility to cognitive biases such as overconfidence has remained unexplored. In this paper, we aim to address this question. Importantly, different from previous research on overconfidence in group judgments that focused solely on a direct comparison between individuals and groups (e.g., Plous 1995, Russo and Schoemaker 1992, Sniezek and Henry 1989), we focus on comparing groups with different gender compositions; that is, groups with different proportions of male and female members. In particular, we explore the link between (a) group gender composition, (b) opinion and information sharing among group members during group deliberations, and (c) group confidence calibration. Moreover, our study strives to provide insights into another question that has remained mostly unclear in prior research (see, e.g., Sniezek 1992, Plous 1995): Under what circumstances is group deliberation a remedy against overconfidence and when might it be ineffective or even exacerbate the problem?

In doing so, our work also provides additional insights into the effects of gender diversity in the upper echelons of organizations. A large number of empirical studies have shown that the inclusion of women into top management teams and boards can have a substantial effect on a variety of organizational outcomes such as financial performance (Post and Byron 2015), risk taking (Baixauli-Soler et al. 2015), and financial fraud (Cumming et al. 2015). Our research complements this prior work by helping to open the "black box" concerning the group processes that link group gender composition and the quality of group decisions.

2. Hypotheses Development

2.1. Confidence Calibration in Individual Judgments

A widely used method to assess confidence calibration is to ask participants for subjective confidence intervals for a number of unknown values. Miscalibration is then defined as the difference between the confidence level and the ratio of the number of times that the true value falls inside of the confidence interval over the total number of questions, where a ratio lower (resp., higher) than the confidence level indicates overconfidence (resp., underconfidence). The most common finding in this paradigm is overconfidence. Even though the degree of observed overconfidence varies depending on the precise nature of the task at hand and the level of confidence according to which individuals are asked to state (e.g., 90% versus 50% or 70%), overconfidence has been demonstrated for estimates in a variety of domains such as general knowledge questions (e.g., Lichtenstein and Fischhoff 1977, Klayman et al. 1999, Soll and Klayman 2004), stock price forecasts (Budescu and Du 2007), or outcomes of sport games (Tsai et al. 2008). Moreover, overconfidence is not limited to estimates made by students in laboratory experiments, but also has been found frequently in judgments by professionals such as financial traders (Glaser et al. 2013), stock market analysts (Deaves et al. 2010, Jain et al. 2013), general managers (Russo and Schoemaker 1992), and chief financial officers (Ben-David et al. 2013). In contrast, systematic underconfidence with respect to confidence calibration has only been very rarely observed (Moore and Healy 2008).

Although the precise relationship between these variables is not perfect, in general, calibration is affected by both judgment accuracy and interval width. For example, decision makers might make relatively accurate judgments, but the estimates could still be badly calibrated if the confidence intervals are set to be very narrow. Conversely, a decision maker could be very inaccurate but still achieve good calibration by setting confidence intervals wide enough.

2.2. Group Deliberation and Confidence Calibration

The general finding from the comparison between *individuals* and *groups* with respect to confidence calibration is that groups are better calibrated than individuals (Plous 1995, Russo and Schoemaker 1992, Sniezek and Henry 1989). Moreover, it appears that such improved calibration in groups compared to individuals is mostly driven by higher accuracy of group judgments rather than groups' greater appreciation of their own limited knowledge. In particular, whereas group judgments in previous studies tended to be more accurate than those of individuals, the confidence intervals set by groups were not wider or in many cases even narrower than those set by individuals (Sniezek 1992, Plous 1995).

In principle, groups have access to a larger and more diverse pool of information than individuals (e.g., Hinsz et al. 1997, Levine and Smith 2013), and group members can exchange arguments in favor of or against a certain position, which should help them to better assess the degree of uncertainty in their judgments (Sniezek and Henry 1989, Sniezek 1992). In particular, disagreements between group members should make group members less confident about their judgments; on the other hand, strong agreement among members should indicate good reason to be confident about an answer. Thus, if all information as well as agreements and disagreements among members are shared openly and in an unbiased manner, one would expect groups to provide better-calibrated confidence statements than individuals. This is also consistent with the explanation for overconfidence in individual judgments by Tversky and Kahneman (1974), who attribute

overconfidence to a process of anchoring on an initial judgment and not adjusting the limits of the confidence interval sufficiently. Following their theory, if group members openly share their opinions and information, a group will have several judgments provided by group members to serve as anchors and thus confidence intervals should be less susceptible to insufficient adjustment.

However, there are several important reasons why groups might not be able to take advantage of the diverse opinions and information present in the group, and thus fail to improve their confidence calibration. First of all, group members are subject to the desire for social acceptance and being liked (Deutsch 1949, Schachter 1959); therefore, they often do not voice dissenting opinions and judgments so as to avoid conflicts (e.g., Asch 1952, Nemeth 1986). Similarly, research on groupthink (e.g., Janis 1982) suggests that the desire to preserve harmony within a group can override the motivation to freely share information, especially when such information contradicts the opinions of other group members. Finally, even in the absence of a desire for social acceptance and group harmony, a group member might simply fail to contribute his or her private information because groups tend to focus their discussion on the information that is already available to all group members before the discussion started (e.g., Stasser 1992).

As a consequence of these processes, group judgments are frequently based on only a small subset of all available opinions and information that could theoretically be shared by group members. This might be particularly harmful for confidence calibration because the influence of a particular group member on group judgment is often strongly linked to his or her individual level of confidence (e.g., Zarnoth and Sniezek 1997). For example, Anderson et al. (2012) demonstrated that overconfident individuals were perceived by other group members as more competent and in turn were awarded higher status and larger influence in the group. This effect even held when group members learned about the overconfident individual's true competence (Kennedy et al. 2013). Therefore, when groups rely only on information and opinions provided by a small subset of group members, those members with the highest degree of individual overconfidence might have the most influence on group judgments and thus drive up group confidence to an unwarranted level.

Whereas a larger degree of opinion and information sharing is likely to cause groups to make better-*calibrated* confidence judgments, compared to groups in which less information is shared, it is less clear whether it will also lead to more *accurate* judgments. In general, group deliberations have the greatest positive effect on judgment accuracy for tasks with solutions that can be easily demonstrated to be correct to others once all information is available, such as mathematical problems (Laughlin and Ellis 1986). In this case, a group member who knows the correct answer can persuade others, and the group will usually perform at the level of its best member or even above (e.g., Laughlin and Ellis 1986, Laughlin et al. 2002). However, for tasks involving estimation of unknown values in which the correct solution cannot be easily demonstrated to others, even if all available information is shared during a group discussion, group members might still not be able to take full advantage of this information and improve accuracy much beyond what would be expected from a simple statistical aggregation of individual judgments. Consistent with this logic, for these estimation tasks, studies found that group judgments with deliberations tend to be more accurate than individual judgments but only similarly accurate as a statistical aggregation of those judgments (e.g., Gigone and Hastie 1997, Sniezek 1990, Tindale and Larson 1992). Consequently, in our studies that focus only on estimation tasks, we do not necessarily expect a significant effect of opinion and information sharing on groups' judgment accuracy.

2.3. The Effects of Gender Composition on Group Deliberations

There is strong evidence that compared to all-male groups, the presence of female group members significantly affects the way group members interact with each other (for an overview, see, e.g., Bear and Woolley 2011). In general, women exhibit higher levels of interpersonal sensitivity—i.e., they pay more attention and show more respect to other people's feelings and thoughts (Fletcher 1998, Hall 1978). Consistently, prior findings have shown that even women in leadership roles tend to be more focused on maintaining a positive relationship with their subordinates compared to male leaders (e.g., Eagly and Johnson 1990). As a consequence of their higher interpersonal sensitivity, women are, for example, less likely than men to obtrusively interrupt others during group discussions (e.g., Anderson and Leaper 1998, Smith-Lovin and Brody 1989) and tend to behave more cooperatively during group tasks than men (Kennedy 2003). In line with these previous findings, compared to all-male groups, groups with female members tend to display more egalitarian behaviors, such as equal communication among group members and shared leadership (Berdahl and Anderson 2005, Mast 2001). Moreover, Woolley et al. (2010) found that a higher share of female group members made group discussions less centered on only a few dominant group members, which enabled all group members to participate more equally in the group discussion. Their results also showed that this effect was strongly linked to female group members' higher level of social sensitivity.

Importantly, in addition to having a direct impact on group interactions because of their own behavior, the presence of female group members also affects the way *male* members behave and interact with other group members (both male and female). Individuals often hold different beliefs and expectations about what constitutes appropriate behavior when interacting with or in the presence of women compared to a setting involving only men (e.g., Williams and Polman 2015). In particular, these beliefs usually involve the importance of showing more interpersonally sensitive behavior in a mixed-gender setting. For example, even today in many settings there is a normative expectation for men to refrain from swearing in the presence of women. Moreover, in addition to changes in behavior due to normative expectations concerning politeness, the presence of female observers or team members has also been found to cause men to behave more generously and helpfully to other group members of both genders (Boschini et al. 2011, Dufwenberg and Muren 2006, Van Vugt and Iredale 2013). Studies conducted outside of the lab have shown similar effects. For example, evidence by Williams and Polman (2015) from teams of management consultants revealed that male consultants in mixed-gender groups were more willing to act with interpersonal sensitivity in interaction with male clients compared to male consultants in all-male groups. Similarly, Adams and Ferreira (2009) found that female directors engaged in more group-oriented behavior by attending board meetings more regularly than men on all-male boards and, importantly, that such behavior also improved the attendance record of male directors.

Altogether, these findings strongly suggest that compared to all-male groups, the presence of women within a group causes a mental shift in all group members (male and female) toward more group-oriented norms. In turn, such a positive group-oriented and psychologically safe atmosphere in groups promotes the sharing of knowledge and the expression of opinions, especially when group members are in disagreement with each other (Edmondson 1999, Hackman 1987, McLeod et al. 1997). In particular, group members in such an environment will be less concerned that voicing disagreements or bringing up new pieces of information might damage social harmony or cause them to be negatively evaluated by others and therefore will focus more on sharing information and opinions. Consistent with this suggestion, interpersonally insensitive behavior such as frequent interruptions has been linked to lower information sharing (e.g., Cooke and Szumal 1994), whereas the opposite behavior, encouraging others to voice their opinions, is associated with higher information sharing (e.g., Leana 1985, Van Dyne and LePine 1998). Similarly, experimental results by Greenhalgh and Chapman (1998) demonstrated that information sharing in dyadic relationships is positively correlated with individuals' perceptions that the other person showed respect, acceptance of one's opinions, and empathy. At the same time, perceptions that the other person displayed behavior interpreted as "being pushy or condescending" has a negative effect on information sharing (Greenhalgh and Chapman 1998).

In summary, we suggest that as a consequence of these above processes, the presence of female members—due to their own behavior and their impact on the behavior of male members—will cause group members to share more opinions and information with each other during the group discussion.

Hypothesis 1. *Group members in groups with at least one female member are more willing to share opinions and information than those in all-male groups.*

Note that we did not specify the exact relationship between the degree to which group members share opinions during the group interaction and the proportion of female members in a group—as long as at least one female member is present. One possible conclusion one might draw from the stream of work mentioned above is that this relationship is strictly positive: group discussion norms become more group oriented as the number of women in a group increases and, consequently, all-female groups would share the most information in this regard, followed by female-majority and male-majority groups. However, as we outlined in our previous discussion, the presence of female group members tends to also shift the behavior of male group members toward higher interpersonal sensitivity. Thus, even the presence of only one female group member in a relatively small group might be sufficient to substantially shift the degree of interpersonally sensitive behavior to a level that is similar to that of all-female groups. In this case, we would expect to find a difference between all-male groups and groups with at least one female member, and we expect information sharing and calibration to be relatively similar across all-female, female-majority, and male-majority groups. As a consequence of this ambiguity concerning the effect of a higher proportion of female members, we refrain from making a prediction concerning the precise relationship between group calibration and proportion of female members. Instead, we focus on the comparison of all-male groups with groups that contain at least one female member.

Because, as discussed above, sharing of opinions and information during the group discussion should have a direct positive effect on groups' confidence calibration, we make the following predictions concerning the effect of the presence of at least one female group member on confidence calibration:

Hypothesis 2A. *Groups with at least one female group member will make better-calibrated confidence judgments than those consisting of only male members.*

Hypothesis 2B. *Better calibration in groups with at least one female member will be mediated by a higher extent of opinion and information sharing during the group deliberation.*

In general, it is possible that better-calibrated judgments in groups with at least one female member are driven by individual differences between men and women in confidence judgments. Existing research found mixed results concerning the effect of gender on miscalibration among individual decision makers. For example, whereas Soll and Klayman (2004) reported that women provided wider confidence intervals than men and were better calibrated, other studies did not find an effect of gender on interval widths or calibration (Biais et al. 2005, Jonsson and Allwood 2003). Importantly, as outlined in the previous discussion, our theoretical predictions do not rely on individual differences in confidence calibration between men and women. Instead, we suggest that independent from such a possible effect, the presence of female group members will strongly affect the extent to which group members share opinions and information with each other and consequently confidence calibration. In addition, we want to highlight that our theoretical predictions strongly built on prior research concerning gender differences in interpersonal sensitivity and thus our hypotheses refer to only the effects of group gender compositions and not those of group diversity in other dimensions such as age or ethnicity (e.g., Van Knippenberg and Schippers 2007, Van Knippenberg et al. 2004).

3. Study 1

3.1. Experiment Design

3.1.1. Methodology. We recruited 352 English-speaking participants (179 male, 173 female; $M_{age} = 23$ years) from a major European university via an online sign-up system. We conducted a total of 14 experimental sessions with approximately 25 participants in each session. Participants were paid a fixed fee of €10. The study had a between-subject design with four group conditions in which we varied group gender compositions: all male ($n = 26$), male majority ($n = 25$), female majority ($n = 23$), and all female ($n = 25$).[1] Participants of corresponding genders were randomly selected into each condition to form groups of three. Participants in these conditions made their judgments after an unstructured face-to-face discussion. All verbal interactions between the three group members were audiotaped with the explicit knowledge of the participants. In addition, we also included two individual conditions in which male ($n = 28$) and female participants ($n = 27$) made their judgments alone without interaction with other participants.

3.1.2. Procedure. Participants were welcomed to the lab and assigned to a computer. In the individual conditions, each participant was seated in front of a computer. In the group conditions, all three group members were seated in front of the same computer. In each group, the group member whose birthday was closest to the date of the experiment was chosen to enter the group judgments into the computer.[2] Participants were not given any particular instructions on how to reach a joint group judgment. Specifically, the instructions were to "use whatever process you like to make your decisions and judgment." The assignment to groups was random except that we made sure that across all sessions there were approximately the same number of groups for each of the four types of group gender composition and that members of any group did not know each other before the study. After participants were seated, they were asked to read the instructions on their screen. Participants were also provided with paper-based versions of the instructions.

We assigned two sets of items to the participants: 10 general-knowledge questions (e.g., Klayman et al. 1999, Russo and Schoemaker 1992, Soll and Klayman 2004) and three forecasting questions (e.g., Budescu and Du 2007, Deaves et al. 2010, Glaser et al. 2013, Jain et al. 2013). Four of those general-knowledge questions asked about distances between cities (Berlin to Vienna, Cairo to Cape Town, Los Angeles to Tokyo, and Paris to Moscow), three about the weights of unknown quantities (an elephant baby born recently in the Zoo of Vienna, an empty Airbus A380, and an empty Opel Astra limousine), and three about the prices of products (an Apple laptop with a number of specific features, an economy-class air ticket from Vienna to New York booked in the next week, and a Mercedes S-Class bought in Austria in the most basic version). Questions were presented in a random order. Participants were asked to state point estimates (their best guesses) as well as upper and lower bounds of 50%, 70%, and 90% confidence interval estimates.

In the first two forecasting questions, participants were asked to provide 90% confidence intervals for the values of the Dow Jones Index and Microsoft share prices in one month, six months, and one year. To not overburden participants, we only asked for 90% intervals for this task. To help with their forecasts, we informed participants of the values of the Dow Jones Index and Microsoft share price at the date of the experiment. The third forecasting task adapted from Jain et al. (2013) asked participants to estimate the future value of a random variable following a random walk. In this task, we provided participants with a description of the change of a random variable over time and asked them to provide 50%, 70%, and 90% confidence intervals for the value of this variable after 100 periods. The initial value of the variable was zero

and, in each period, there was an equal chance of either a one-unit increase or a one-unit decrease. Hence, the expected value of this variable is zero and its variance is simply the total number of periods $T = 100$. The theoretical 50%, 70%, and 90% confidence intervals for this variable are, respectively, $(-0.674\sqrt{T}, 0.674\sqrt{T})$, $(-1.036\sqrt{T}, 1.036\sqrt{T})$, and $(-1.645\sqrt{T}, 1.645\sqrt{T})$.

After finishing the 10 general-knowledge and the three forecasting questions, all participants were asked to individually fill in a final paper-based questionnaire with demographic information. In addition, participants in the group conditions answered questions that aimed to assess group members' satisfaction with their groups and the extent to which group members shared opinions and information during the discussion. At the end of the questionnaire, all participants were asked what they believed to be the purpose of the experiment. Only four participants correctly guessed that gender was a factor in our study. Removing these participants from the sample did not change any of our results.

3.1.3. Measures. Based on participants' point and interval estimates, we composed several measures that capture the quality of their judgments. In the following, we introduce our notations and the specific measures. Let X_i denote the quantity that question i asks for (e.g., the weight of an empty A380 or the price of Microsoft shares in six months). Since the participants do not know the answer for sure, X_i is a random variable to them, and its realization, denoted as x_i, stands for either the correct answer to a general-knowledge question or the actual value of a forecasted item. We let m_{ij} denote a decision maker (either an individual or a group) j's point estimate for question i. Moreover, l_{ijk} and u_{ijk} denote, respectively, the lower bound and upper bound of decision maker j's confidence interval estimate at confidence level k to question i.

Judgment calibration. For the general-knowledge questions, we used both *hit rate* and *calibration error* to measure the calibration of participants' estimates. The hit rate for a decision maker j at confidence level k was computed by counting the number of times the true value was within the confidence interval across the 10 questions divided by 10, $h_{jk} = (1/10)\sum_{i=1}^{10} \mathbf{1}_{l_{ijk} \le x_i \le u_{ijk}}$, and the calibration error was calculated as the absolute difference between the hit rate and the required confidence level, $e_{jk} = |h_{jk} - k|$; $i = 1, 2, \ldots, 10$, $j = 1, 2, \ldots, n$, n being the total number of decision makers in a given condition, k={50%, 70%, 90% }, and $\mathbf{1}_A$ an indicator function that equals 1 if the condition A is satisfied and 0 otherwise.[3] A decision maker j is considered to be perfectly calibrated, underconfident, or overconfident when the hit rate h_{jk} for a given confidence level k equals, is greater than, or is less than the corresponding confidence level k, respectively, and the calibration error captures the decision maker's degree of miscalibration—the larger the calibration error, the less calibrated is the decision maker's judgment.

For the two financial forecast questions, since we only have data for two forecasts, it is not meaningful to use the hit rate as a normative benchmark. Therefore, following the prior literature on stock-market forecasts (e.g., Ben-David et al. 2013, Glaser et al. 2013) and consistent with the definition of overconfidence as an overestimation of signal precision (e.g., Odean 1998), we derived *return volatility* estimates from the decision makers' confidence interval estimates and then used the mean historical return volatility of the Dow Jones Index and Microsoft share price as a normative benchmark. To do that, for each item (either the index value or the share price), we first transformed decision maker j's stated confidence intervals into intervals for returns by dividing the interval's upper and lower bounds by the corresponding item's value on the day of the experiment: $u_j' = u_j/x_0$ and $l_j' = l_j/x_0$. We then deduced decision maker j's implicit volatility estimate using the following approximation (Pearson and Tukey 1965, Keefer and Bodily 1983): $v_j = (u_j' - l_j')/3.25$. A decision maker j is considered to be perfectly calibrated, underconfident, or overconfident when the estimated return volatility v_j is equal to, greater than, or less than the mean historical return volatility, v_0, which we calculated using stock-price data obtained from the Center for Research in Security Prices (CRSP) ranging from 1995 to 2015.

Judgment accuracy. Our accuracy measure for a general-knowledge question i by decision maker j is the *absolute percentage error* computed by taking the absolute difference between a point estimate and the true value, divided by the true value (see, e.g., Mannes et al. 2014, Minson and Mueller 2012, Davis-Stober et al. 2014): $a_{ij} = |m_{ij} - x_i|/x_i$.

Confidence interval width. We computed a measure of confidence interval width to capture the extent to which participants appreciated uncertainty around their point estimates. The *percentage interval width* of question i by a decision maker j at confidence level k was calculated as $w_{ijk} = (u_{ijk} - l_{ijk})/m_{ij}$.

Opinion and information sharing. To measure the level of *opinion and information sharing* during the group deliberations, we asked each participant in the group conditions to rate four items adapted from Phillips and Loyd (2006) on a 1 = "not at all" to 7 = "very much" scale: (i) "Group members listened to each other's point of view," (ii) "Group members encouraged each other to share their opinions," (iii) "Group members were interested in what theothers had to say," and

(iv) "Group members shared a lot of information with each other."

Group member satisfaction. We measured *group member satisfaction* with three items adapted from Jehn et al. (2010) on a 1 = "not at all" to 7 = "very much" scale: (a) "I was very satisfied working in this group during this exercise," (b) "I would like to work with this group again," and (c) "I was happy working in this group during this exercise."

3.2. Results

3.2.1. Results from General-Knowledge Questions.

For our main measures of interest, we initially also tested for differences between judgment categories (distances, weight, and price) but did not find significant main effects or interactions with our experimental conditions. Therefore, we dropped this variable from the analysis. We also tested for the effect of diversity with respect to age and ethnicity, which together with gender have been demonstrated to be the most important dimensions of demographic diversity in small groups (e.g., Mannix and Neale 2005, Van Knippenberg and Schippers 2007). Since 93% of our participants were white and 90% of them were between the age of 21 and 27 (total range: 18–32), groups were in general very homogenous with respect to these two factors. Our analysis of age diversity or the presence of nonwhite group members indeed showed no significant effect on any of our dependent measures. Hence, to focus on our main results, we do not discuss these two factors further.

Judgment calibration. Table 1 presents hit rates and calibration errors for the 50%, 70%, and 90% confidence intervals over all 10 questions across the six different types of decision makers.

Our results show that for the 50% confidence level, 30% of decision makers exhibited underconfidence, 16% were perfectly calibrated, and 55% were overconfident. For the 70% confidence level, the proportions of underconfident, perfectly calibrated, and overconfident decision makers were 10%, 19%, and 71%, respectively. Finally, for the 90% confidence level, 1% of decision makers were underconfident, 7% were perfectly calibrated, and 92% were overconfident.

To analyze the effect of decision-maker type on miscalibration, we conducted a 6 (decision-maker type) × 3 (confidence level) mixed ANOVA of calibration errors. The results showed a significant main effect of confidence level, $F(2,296) = 81.82$,[4] $p < 0.01$, a significant main effect of decision-maker type, $F(5,148) = 4.84$, $p < 0.01$, and a significant interaction effect, $F(10,296) = 4.55$, $p < 0.01$. Providing support for Hypothesis 2A, our follow-up analysis with planned contrasts revealed that tested jointly across all three confidence levels,[5] all-male groups were significantly worse calibrated than groups of other gender compositions, $F(3,296) = 11.68$, $p < 0.01$, $d = 0.56$. Tested for each probability level separately, this effect was significant for the 90%, $F(1,296) = 25.59$, $p < 0.01$, and 70% confidence level, $F(1,296) = 9.42$, $p < 0.01$, but not significant for the 50% confidence level, $F(1,296) = 0.04$, $p = 0.85$. Furthermore, pairwise comparisons showed that there was no significant difference in calibration errors among groups with a majority of male members, a majority of female members, or female members only, p's > 0.25.

Moreover, we found that, jointly tested across all three confidence levels, there was no significant difference in calibration errors between individual judgments by men and women, $F(3,296) = 1.40$, $p = 0.25$. When we tested for each confidence level separately, we found a marginally significant difference at the 90% level, $F(1,296) = 3.32$, $p = 0.07$, but not at others, p's > 0.39. Furthermore, our results showed that, tested jointly across all three confidence levels, groups with at least one female member were significantly better calibrated than male, $F(3,296) = 21.73$, $p < 0.01$, $d = 0.75$, or female, $F(3,296) = 35.75$, $p < 0.01$, $d = 0.97$, individuals; however, whereas all-male groups were still significantly better calibrated than female individuals (though with a smaller effect size than other groups, $F(3,296) = 4.32$, $p = 0.01$, $d = 0.40$), there was

Table 1. Hit Rates and Calibration Errors Across Confidence Levels and Decision-Maker Types (Study 1)

		Hit rate (%)						Calibration error (%)					
		50%		70%		90%		50%		70%		90%	
	Decision-maker type	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Individual	Gender												
	Female	39.26	13.57	46.67	13.87	54.81	15.03	14.44	9.34	24.07	12.48	35.19	15.03
	Male	41.43	13.80	50.00	15.63	58.93	14.74	13.57	8.70	22.14	12.28	31.07	14.74
Group	Gender composition												
	All female	49.60	18.37	61.20	15.63	70.40	12.07	15.60	9.17	13.60	11.50	20.40	10.60
	Female majority	50.87	14.43	62.61	14.84	70.87	13.11	11.30	8.69	11.74	11.54	19.13	13.11
	Male majority	48.80	15.63	61.20	17.40	72.80	16.46	12.40	9.26	13.60	13.81	18.00	15.55
	All male	40.38	13.99	51.15	13.95	61.15	12.43	13.46	10.18	18.85	13.95	28.85	12.43

no significant difference between all-male groups and male individuals, $F(3, 296) = 1.02$, $p = 0.39$.

We next compared group judgments with judgments that result from a simple statistical aggregation procedure of judgments by three randomly determined individual decision makers. The main goal of this analysis was to explore the extent to which our results could be explained by a simple aggregation of individual judgments (e.g., Gigone and Hastie 1997). In particular, it is at least theoretically possible that although we do not find a significant effect of gender on calibration for individual judgments, the aggregation of judgments by male or female individuals might still result in a pattern that is similar to the one that we have observed for interacting groups. Moreover, such a comparison also has practical implications as it enables us to assess the extent to which organizations would be better or worse off by simply aggregating individual judgments—as opposed to forming a team of individuals and achieving a consensus judgment through face-to-face interactions (Gigone and Hastie 1997).

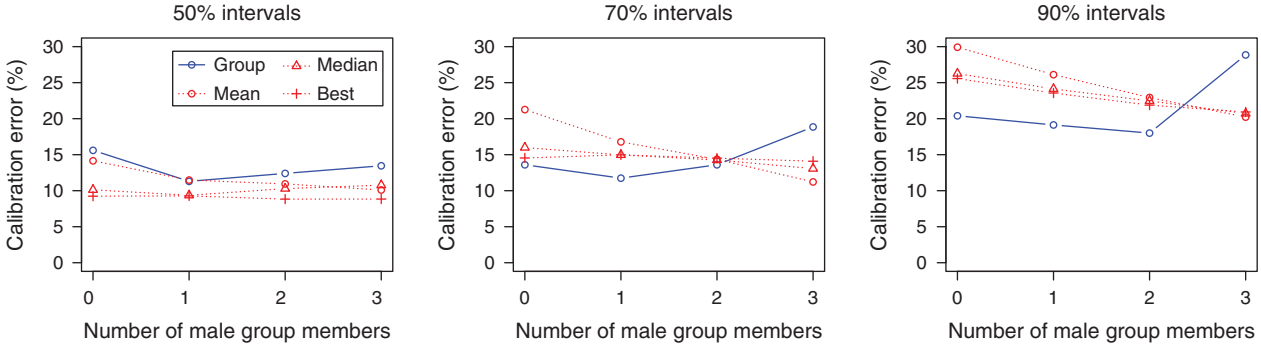
Based on prior research on the aggregation of expert judgments (e.g., Mannes et al. 2014), we considered three different aggregation models: aggregation by taking the (a) mean or (b) median of three randomly selected individual judgements, or by (c) determining the top performer out of three individual decision makers and implementing only the judgments of the best member. To compute calibration errors that would be expected from such an aggregation, we repeatedly and randomly sampled three individuals from the individual conditions and implemented the three aggregation procedures for each sample. To implement the mean or median model, for each question, we composed the aggregated confidence interval by taking the mean or median of the three lower and upper bounds stated by the three sampled individuals and then computed the calibration error based on the aggregated confidence intervals over the 10 questions. To implement the best-member model, out of the three sampled individuals, we selected the individual with the lowest calibration error computed over all 10 questions and then computed calibration errors based on the judgments of this individual. We repeated this process 1,000 times (sampling with replacement) and then averaged the calibration errors across the 1,000 trials (see, e.g., Gaba et al. 2017, Hora 2004, and Park and Budescu 2015 for a similar methodology). To ensure that our results are comparable with the group estimates, we conducted this process for each gender composition separately where the gender of the three selected individuals was always consistent with the corresponding gender composition of the group. Specifically, we aggregated intervals estimated by (i) three female individuals, (ii) two female individuals and one male individual, (iii) two male individuals and one female individual, or (iv) three male individuals, and then compared them with those estimated by all-female, female-majority, male-majority, and all-male interacting groups, respectively. Figure 1 shows the calibration errors from the interacting groups and the three aggregation procedures.

Most of the miscalibration in the aggregation models resulted from overconfidence. In particular, the proportion of overconfident outcomes from the aggregation procedure for the 50%, 70%, and 90% intervals, respectively, was 52%, 72%, and 93% for the mean model, 38%, 72%, and 95% for the median model, and 48%, 84%, and 100% for the best-member model.

As Figure 1 shows, for the 70% and 90% confidence levels, groups with at least one female member were better calibrated than what would be expected from a simple mean aggregation of confidence intervals, whereas the difference was only very small for the 50% confidence level. Averaged over the three confidence levels, this difference was significant for all-female, $t(24) = 3.15$, $p < 0.01$, and female-majority groups, $t(22) = 2.31$, $p = 0.03$, but not for male-majority groups, $t(24) = 0.65$, $p = 0.53$. Interestingly, calibration for all-male groups was significantly worse than the outcome of the mean model, $t(25) = 3.39$, $p < 0.01$.

Similarly, we find that all-male groups are also significantly worse calibrated compared to the outcome

Figure 1. (Color online) Calibration Errors from Group Estimates and Aggregated Individual Estimates (Study 1)



of the median model, $t(25) = 2.82$, $p = 0.01$. For all other group types, there was no significant difference between the aggregation model and the group results, p's > 0.68. Finally, comparing the group results to those from the best-member model, we again find that all-male groups were significantly worse calibrated than the outcome of the aggregation model, $t(25) = 2.98$, $p = 0.01$, and that there was no significant difference for any of the other group types, p's > 0.49.

Judgment accuracy and confidence interval widths. There are two important factors that might affect the differences in calibration errors across decision makers: accuracy and interval widths. We next test the extent to which these two factors can at least partially account for our observed difference in calibration errors between all-male groups and other groups.[6]

Table 2 presents the absolute percentage errors as well as the 50%, 70%, and 90% percentage interval widths averaged over all 10 questions across the six decision-maker types.

A one-way ANOVA of absolute percentage errors averaged over all 10 questions showed a significant effect of decision-maker type, $F(5, 148) = 3.97$, $p < 0.01$. Planned contrasts did not show a significant difference between all-male groups and the other groups, $F(1, 148) = 0.00$, $p = 0.98$, or between individual judgments by men and women, $F(1, 148) = 1.36$, $p = 0.25$. In contrast, our results did show that groups made significantly more accurate judgment than individuals, $F(1, 148) = 17.47$, $p < 0.01$, $d = 0.71$.

A 6 (decision-maker type) × 3 (confidence level) mixed ANOVA of percentage interval widths averaged over all 10 questions revealed a significant main effect of decision-maker type, $F(5, 148) = 3.04$, $p = 0.01$, a significant main effect of confidence level, $F(2, 296) = 509.05$, $p < 0.01$, and a significant interaction effect between the two factors, $F(10, 296) = 3.68$, $p < 0.01$. Our follow up analysis with planned contrasts tested jointly across all three confidence levels revealed that the percentage interval widths of all-male groups were significantly smaller than those of other groups, $F(3, 296) = 22.62$, $p < 0.01$, $d = 0.56$, male individuals, $F(3, 296) = 18.54$, $p < 0.01$, $d = 0.73$, or female individuals, $F(3, 296) = 15.32$, $p < 0.01$, $d = 0.74$. For individual judgements, there was no significant difference between judgments by women and men, $F(3, 296) = 0.52$, $p = 0.69$. Finally, we found that the interval widths provided by groups with one or more female members were not significantly different from those provided by female, $F(3, 296) = 0.14$, $p = 0.94$, or male, $F(3, 296) = 0.30$, $p = 0.83$, individuals.

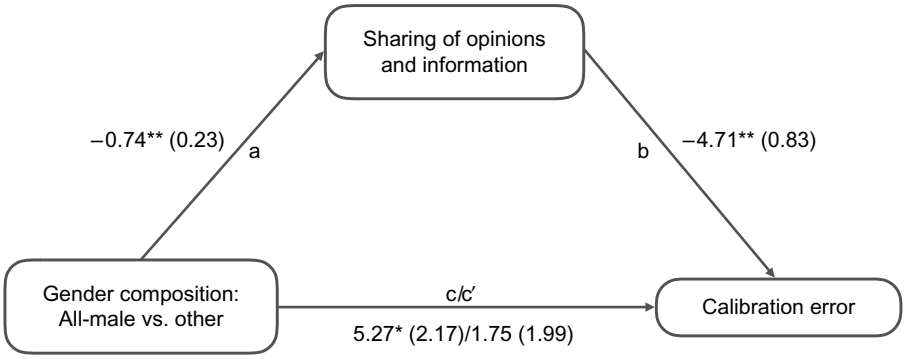
Mediation and group discussion. We averaged the four items measuring the degree of opinion and information sharing during the group discussion into one composite measure ($\alpha = 0.86$).[7] Interrater reliability across the three group members (ICC[1] = 0.61) was significantly different from zero, $F(98, 198) = 5.66$, $p < 0.01$, suggesting that group members' ratings were strongly interdependent. Based on this result, we then further aggregated the three composite measures of the group members into one group measure (e.g., Kozlowski and Klein 2000). A one-way ANOVA across the four gender compositions showed a significant effect of gender composition on the degree of opinion and information sharing in groups, $F(3, 95) = 3.75$, $p = 0.01$. In particular, as suggested by Hypothesis 1, all-male groups engaged significantly less in the exchange of opinions and information than other groups, $F(1, 95) = 10.53$, $p < 0.01$, $d = 0.74$. There was no significant difference between any of the other group types, p's > 0.40.

We next tested whether the degree of opinion and information sharing during the group discussion mediated the difference in calibration errors between all-male groups and other group types, as suggested in Hypothesis 2B. Figure 2 presents the results of a mediation analysis (Baron and Kenny 1986) with opinion and information sharing as the mediator between group gender composition (all-male groups versus other group types) and calibration error averaged over all three confidence levels.

Table 2. Absolute Percentage Errors and Percentage Interval Widths Across Decision-Maker Types (Study 1)

		Absolute percentage error		Percentage interval width					
				50%		70%		90%	
	Decision-maker type	M	SD	M	SD	M	SD	M	SD
Individual	Gender								
	Female	80.58	46.32	70.43	13.90	103.20	22.34	140.33	38.87
	Male	92.80	53.43	75.19	16.45	106.47	29.34	141.40	47.46
Group	Gender composition								
	All female	64.05	26.43	73.27	15.49	108.04	24.94	148.16	39.28
	Female majority	59.85	42.06	72.02	14.09	98.24	14.78	125.38	17.75
	Male majority	54.12	22.40	72.93	10.64	106.13	19.95	146.46	41.05
	All male	59.55	29.93	66.45	13.14	93.86	28.30	109.50	19.28

Figure 2. Results of Mediation Analysis (Study 1)



Notes. OLS regression coefficients. Standard errors in parentheses; $^{*}p < 0.05$; $^{**}p < 0.01$.

Our results show that sharing of opinions and information was significantly lower for all-male groups (path a) than for groups with at least one female member. Furthermore, the positive effect of all-male gender composition on calibration error (path c) was reduced—and actually became insignificant—when we controlled for opinion and information sharing in the regression (path c and c′). We next used a bootstrap procedure (Shrout and Bolger 2002) with 5,000 trials to construct a 95% confidence interval for the indirect effect of gender composition on calibration error. The confidence interval (0.84, 6.20) excluded zero, indicating that our measure of opinion and information sharing is a significant mediator.

3.2.2. Results from Financial Forecasts and Random-Walk Forecasts

Financial forecasts. A comparison of return volatility estimates for the Dow Jones Index and its mean historical return volatility (9.83%), both averaged over all three time horizons, revealed that both groups (M = 5.38%, SD = 3.93%), $t(98) = 11.18$, $p < 0.01$, and individuals (M = 4.75%, SD = 3.77%), $t(54) = 9.93$, $p < 0.01$, displayed overconfidence and significantly underestimated return volatilities. For Microsoft shares[8] (mean historical return volatility 25.58%), we found that individuals (M = 30.11%, SD = 17.09%) actually overestimated return volatilities, $t(53) = 3.85$, $p < 0.01$, whereas groups (M = 20.85%, SD = 12.81%) underestimated return volatilities (256), $t(96) = 1.95$, $p = 0.06$.[9]

We next conducted two separate 6 (decision-maker type) × 3 (time horizon) mixed ANOVAs of return volatility estimates for the Dow Jones Index and the Microsoft share price. Our analysis for the Dow Jones Index revealed a significant effect of time horizon, $F(2, 296) = 113.50$, $p < 0.01$, and decision-maker type, $F(5, 148) = 2.61$, $p = 0.03$; the interaction between the two was marginally significant, $F(10, 296) = 1.76$, $p = 0.07$. Planned contrasts tested jointly across all three time horizons showed that return volatility estimates by all-male groups (M = 38.07%, SD = 29.70%) were significantly lower than those by groups with at least one female member (M = 59.40%, SD = 40.95%), $F(3, 296) = 11.56$, $p < 0.01$, $d = 0.56$. We also found that return volatilities by groups with at least one female group member were (marginally) significantly higher than those by male (M = 45.90%, SD = 31.57%), $F(3, 296) = 4.47$, $p < 0.01$, $d = 0.35$, and females (M = 49.25%, SD = 43.69%), $F(3, 296) = 2.50$, $p = 0.06$, individual. In contrast, return volatilities stated by all-male groups were marginally significantly lower than those by female individuals, $F(3, 296) = 2.35$, $p = 0.07$, and not significantly different from those estimated by male individuals, $F(3, 296) = 1.33$, $p = 0.27$. There was no significant difference between return volatility estimates by individual men and women, $F(3, 444) = 0.20$, $p = 0.90$.

Similarly, for estimated return volatilities of Microsoft shares, the results showed a significant effect of time horizon, $F(2, 293) = 128.87$, $p < 0.01$, and of decision-maker type, $F(5, 148) = 4.55$, $p < 0.01$, but there was no significant interaction effect, $F(10, 293) = 0.87$, $p = 0.57$. Planned contrasts tested jointly across all time horizons revealed that return volatility estimates by all-male groups (M = 14.07%, SD = 9.65%) were significantly lower than those by other groups (M = 23.32%, SD = 12.98%), $F(3, 293) = 14.08$, $p < 0.01$, $d = 0.73$. There was no significant difference between return volatility estimates by individual men (M = 28.80%, SD = 16.03%) and women (M = 31.53%, SD = 18.34%), $F(3, 293) = 0.60$, $p = 0.62$. Moreover, the results showed that return volatility estimates by groups with at least one female member were significantly lower than those by female, $F(3, 293) = 8.97$, $p < 0.01$, $d = 0.51$, or male, $F(3, 296) = 5.44$, $p < 0.01$, $d = 0.42$, individuals. Moreover, return volatilities by all-male groups were also significantly lower than those by male, $F(3, 293) = 24.26$, $p < 0.01$, $d = 1.10$, or female, $F(3, 293) = 30.55$, $p < 0.01$, $d = 1.09$, individuals.

Random-walk forecasts. Similar to the financial forecasts, it is not meaningful to use the hit rate as a normative benchmark since we only have data for one forecast; furthermore, unlike the Dow Jones Index

or the Microsoft share price, there is no meaningful and observable realization of the random variable. On the other hand, the random-walk model allows us to directly compute theoretical confidence intervals that can be used as a benchmark (interval width 22.36 averaged over the three confidence levels). Hence, we are going to compare normative confidence interval widths with those stated by decision makers to assess the susceptibility that a decision maker is prone to miscalibration. Unlike in our previous analyses, we do not use the percentage interval width here, as the theoretical expected value of the random walk equals zero.

Our results showed that whereas the interval widths of groups (M = 25.87, SD = 21.21) was not significantly different from the normatively correct value, $t(98) = 1.65$, $p = 0.10$, intervals provided by individuals (M = 13.40, SD = 12.01) were significantly too narrow, $t(54) = 5.53$, $p < 0.01$.

A 6 (decision-maker type) × 3 (confidence level) mixed ANOVA of confidence interval widths revealed a significant main effect of confidence level, $F(2, 296) = 139.66$, $p < 0.01$, and decision-maker type, $F(5, 148) = 4.92$, $p < 0.01$, and a significant interaction, $F(10, 296) = 4.37$, $p < 0.01$. Confidence intervals by groups were significantly wider than those by individuals, $F(3, 296) = 51.04$, $p < 0.01$, $d = 0.67$. In contrast, there was no significant difference in confidence interval widths between all-male groups (M = 25.69, SD = 16.62) and other group types (M = 25.92, SD = 22.73), $F(3, 296) = 0.30$, $p = 0.82$.[10]

3.2.3. Analysis of Audiotapes and Reported Satisfaction. We now turn to our analysis of the audio recordings of group discussions and group members' self-reported satisfaction. In the following, we focus on comparing all-male groups with other group types. Interested readers are referred to Online Appendix Table A5 for a complete summary of all measures across all four group compositions. For all of our survey- and audiotape-based measures, we also tested for differences among the other three group types but did not find systematic significant differences.

On average, group discussions lasted for 32 minutes (SD = 5.64). A one-way ANOVA indicated a significant difference in the amount of discussion time across group types, $F(3, 95) = 3.16$, $p = 0.03$, and a direct comparison showed that discussions in all-male groups (M = 29.04, SD = 6.74) were significantly shorter than those in other groups (M = 32.57, SD = 4.91), $F(1, 95) = 8.14$, $p = 0.01$, $d = 0.65$.

To analyze the extent to which discussions in all-male groups were dominated by only one or two group members, we measured the proportional amount of time each group member was speaking during the discussion and used this measure as a proxy for each group member's participation intensity (e.g., Phillips and Loyd 2006, Woolley et al. 2010, Tost et al. 2013). We then computed the variance of group member participation intensity across the three group members (e.g., Woolley et al. 2010). The participation variance would equal zero when all group members participated equally in the discussion and reach its maximum when only one group member spoke and the remaining two members remained completely silent. A one-way ANOVA revealed a significant difference in the participation variance across groups of different gender compositions, $F(3, 95) = 2.87$, $p = 0.04$. In addition, our analysis showed that the variance in group members' participation intensity during the discussion was significantly larger in all-male groups (M = 0.046, SD = 0.033) than in all other groups (M = 0.029, SD = 0.031), $F(1, 95) = 7.50$, $p = 0.01$, $d = 0.63$.

To analyze participants' satisfaction with their groups, for each group, we aggregated the three self-reported measures of group member satisfaction into one composite variable ($\alpha = 0.81$) and computed the group average. A one-way ANOVA indicated a significant difference across group gender compositions, $F(3, 95) = 2.80$, $p = 0.04$. Moreover, planned contrasts revealed that group members in all-male groups (M = 4.34, SD = 0.61) were less satisfied than those in groups with at least one female member (M = 4.75, SD = 0.79), $F(1, 95) = 5.69$, $p = 0.02$, $d = 0.54$.

Summary. Supporting Hypothesis 2A, for both general-knowledge questions and financial forecasts, we found that all-male groups were significantly worse calibrated than groups of other gender compositions. In contrast, there was no significant difference in calibration errors among male-majority, female-majority, or all-female groups. We further found that this difference between all-male groups and other group types was not driven by individual differences, nor can it be explained by a simple aggregation of individual judgments. Rather, supporting Hypotheses 1 and 2B, differences in calibration errors were driven by lower information sharing in all-male groups. Consistently, our analysis of the recorded group discussions showed that in all-male groups, group interactions were shorter and characterized by more unequal participation patterns.

4. Study 2

Study 2 had a very similar procedure and design as in Study 1. In particular, we employed the same six between-subject conditions: four group conditions with varying gender compositions and two individual conditions with male or female participants. The main goal of the new study was to replicate our previous findings with higher statistical power than before[11] and to address several drawbacks and limitations of Study 1. In particular, in Study 1, we chose the 10 general-knowledge questions arbitrarily from items used in previous research. Although this is a common

approach to assessing overconfidence, other work has argued that overconfidence in general might at least partially be an artifact of a biased question-selection procedure (e.g., Gigerenzer et al. 1991, Juslin et al. 2000). To avoid this potential problem, we now focused on only two specific knowledge domains with which our participants are generally familiar, and within these two domains, we randomly selected questions to compose a representative question set (Gigerenzer et al. 1991). Moreover, one other limitation of Study 1 was that participants' compensation was not explicitly linked to their performance in the actual task. This could in principle affect our results; for example, prior findings suggest that women react differently to competitive financial incentives than men (Gneezy et al. 2003). Although our tasks are not competitive in nature, and thus there is no direct reason to assume that financial incentives should interact with our manipulation, testing for the robustness of our results when financial incentives are used is nevertheless desirable. Therefore, in Study 2, we employed an incentive scheme adapted from Jose and Winkler (2009) that incentivizes participants to report carefully considered intervals. Finally, we introduced two additional new measures that directly measure interpersonally sensitive behavior by male and female group members during the group discussion.

4.1. Experimental Design

4.1.1. Methodology and Procedure. Study 2 followed the same general procedure as Study 1. We recruited 494 (249 male, 245 female; $M_{age} = 24$ years) English-speaking students from a large Austrian university. Participants received on average €12 for their participation. In total, we conducted 17 experimental sessions with approximately 25 participants in each session; each session lasted approximately 50 minutes. Similar to Study 1, participants were assigned to either the two individual conditions (33 female[12] and 34 male participants) or one of the four group conditions: all male ($n = 36$), male majority ($n = 36$), female majority ($n = 35$), and all female ($n = 35$). For the 10 general-knowledge questions, participants were asked to provide point estimates as well as upper and lower bounds of 50%, 70%, and 90% confidence intervals. We selected the 10 general-knowledge questions randomly from two knowledge domains that are at least moderately familiar to our participants.[13] In particular, we created five questions each from the random selection of (a) five pairs of European Union capitals (out of 378 possible pairs) and (b) five electronic products from the university's online shop (out of 63 total items). The 10 selected questions were the distances between (a1) Sofia and Madrid, (a2) Valletta and Stockholm, (a3) Riga and Ljubljana, (a4) Valletta and Dublin, and (a5) Rome and Helsinki, and the prices of (b1) an Apple iPad Air 2, (b2) a Lenovo ThinkPad, (b3) an Apple MacBook Air, (b4) a Microsoft Surface Book, and (b5) a Lenovo ThinkPad X1 Yoga. Participants were informed about the corresponding countries of the given cities and about the technical features of the given electronic products that were listed on the university shop's website. In the financial forecasting questions, participants were asked to provide point estimates as well as 90% confidence intervals for the value of the German stock market index (DAX) in 1, 6, and 12 months' time.[14]

To encourage participants to consider their answers carefully, we employed an incentive scheme adapted from Jose and Winkler (2009). In particular participants were informed that, in addition to a fixed fee of €8, their final payoff would depend on the quality of their judgments, with up to 6€ of additional compensation. They then received a careful explanation of the precise procedure from which their payoffs would be calculated, with an emphasized highlight that their financial payoffs are maximized when they put effort into the task and state estimates that best reflect their actual beliefs (see Jose and Winkler 2009 for a detailed description of the incentive scheme).

4.1.2. Measures. We employed the same measures as in Study 1 to analyze *hit rate*, *calibration error*, *absolute percentage error*, *percentage interval width* for general-knowledge questions, *return volatility estimate* for financial forecasts and *opinion and information sharing*, and *discussion length, participation variance, group member satisfaction* for group discussions. In addition, we included two new measures: *interruption* and *encouragement to participate*. Frequent interruptions are a sign of interpersonally insensitive behavior and an indicator of dysfunctional communication patterns that is likely to decrease information sharing (e.g., Cooke and Szumal 1994). Contrarily, encouraging others to speak is an example of interpersonally sensitive behavior that might increase the exchange of information (Leana 1985, Van Dyne and LePine 1998). Specifically, on completion of the main tasks, participants were asked to rate two types of behaviors by each of the other two group members on a scale from 1 = "not at all" to 7 = "very much": (a) "to what extent were you interrupted by this group member during the group discussion?" and (b) "to what extent did this group member encourage you to participate in the group discussion?".

4.2. Results

Like in Study 1, we initially also tested for effects of knowledge domains (distances and prices) and of group age and ethnicity composition. Again, we did not find a systematic influence of these factors on our main measures of interest. Groups were again quite homogenous with respect to age and ethnicity: the age range was 18–28, and 89% of all group members identified themselves as white.

Table 3. Hit Rates and Calibration Errors Across Confidence Levels and Decision-Maker Types (Study 2)

		Hit rate (%)						Calibration error (%)					
		50%		70%		90%		50%		70%		90%	
	Decision-maker type	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Individual	Gender												
	Female	27.88	19.49	36.67	16.52	49.39	19.68	26.36	12.95	33.33	16.52	40.61	19.68
	Male	27.65	18.10	42.06	20.12	52.65	23.78	25.29	13.54	29.71	17.32	37.35	23.78
Group	Gender composition												
	All female	36.57	17.81	47.71	17.34	61.71	16.89	18.00	13.02	22.86	16.55	28.29	16.89
	Female majority	33.71	16.99	47.43	17.71	59.14	15.79	20.29	11.75	24.29	15.20	30.86	15.79
	Male majority	35.28	15.21	49.17	17.13	63.61	15.52	16.94	12.61	23.06	13.90	26.94	14.51
	All male	24.44	14.03	36.94	14.51	49.44	16.20	26.11	12.93	33.06	14.51	40.56	16.20

4.2.1. Results from General-Knowledge Questions

Judgment calibration. Table 3 presents hit rates and calibration errors for the 50%, 70%, and 90% confidence intervals averaged over all 10 questions across the six types of decision makers.

As shown in Table 3, we find overconfidence across all decision-maker types and confidence levels. For the 50% (resp., 70%, 90%) confidence level, 10% (resp., 4%, 1%) of decision makers are underconfident, 11% (resp., 8%, 3%) perfectly calibrated, and 79% (resp., 88% and 97%) overconfident.

The results of a 6 (decision-maker type) × 3 (confidence level) mixed ANOVA of calibration errors revealed a significant main effect of confidence level, $F(2,406) = 73.46$, $p < 0.01$, and decision-maker type, $F(5,203) = 5.12$, $p < 0.01$, but no significant interaction effect is observed, $F(10,406) = 0.45$, $p = 0.92$. Lending support to Hypothesis 2A, a planned contrast tested jointly across all three confidence levels showed that the calibration of all-male groups was significantly worse than that of other groups, $F(3,406) = 25.91$, $p < 0.01$, $d = 0.80$. This result also held when we tested for differences separately for the 90% confidence level, $F(1,406) = 37.29$, $p < 0.01$, the 70% confidence level, $F(1,406) = 24.72$, $p < 0.01$, and the 50% confidence level, $F(1,406) = 15.72$, $p < 0.01$. On the other hand, our analysis did not reveal a significant difference in calibration errors among groups with at least one female member, p's > 0.18.

We also did not find a significant difference in calibration errors between individual men and women, $F(3,406) = 1.37$, $p = 0.25$. In contrast, we found that groups with at least one female member were significantly better calibrated than female, $F(3,406) = 25.14$, $p < 0.01$, $d = 0.78$, or male, $F(3,406) = 13.72$, $p < 0.01$, $d = 0.55$, individuals. However, this was not the case when we compared all-male groups with female, $F(3,406) = 0.01$, $p = 0.99$, or male, $F(3,406) = 1.27$, $p = 0.28$, individuals.

Like in Study 1, we next aggregated individual confidence intervals by repeatedly and randomly selecting three individual judgments and aggregating them using the mean, median, or best-member models. Figure 3 shows the results from the aggregation procedure.

Similar to interacting groups, miscalibration from the aggregation models mostly resulted from overconfidence. Specifically, the proportion of overconfident outcomes from the aggregation procedure for the 50%, 70%, and 90% intervals was, respectively, 75%, 82%, and 87% for the mean model, 70%, 83%, and 90% for the median model, and 61%, 80%, and 97% for the best-member model.

For groups with at least one female member, t-tests for calibration errors averaged over all three confidence levels showed no significant difference between group

Figure 3. (Color online) Calibration Errors from Group Estimates and Aggregated Individual Estimate (Study 2)

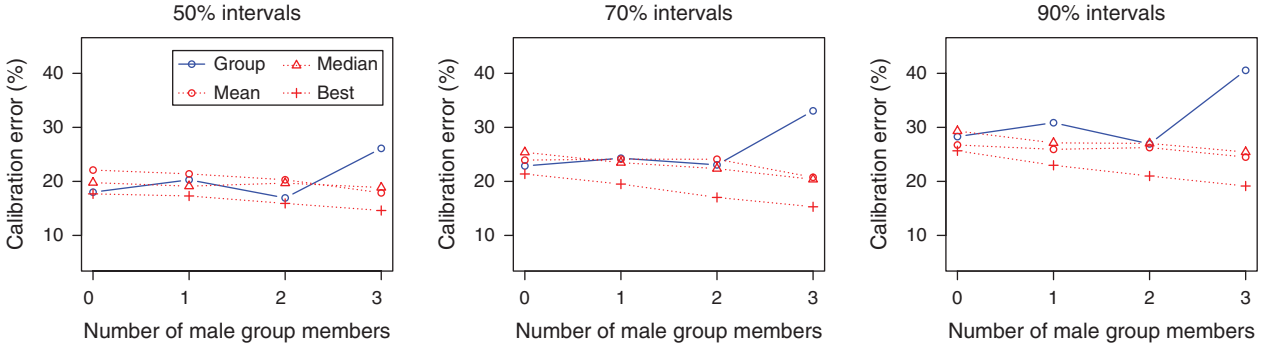


Table 4. Absolute Percentage Errors and Percentage Interval Widths Across Decision-Maker Types (Study 2)

		Absolute percentage error		Percentage interval width					
				50%		70%		90%	
	Decision-maker type	M	SD	M	SD	M	SD	M	SD
Individual	Gender								
	Female	37.82	16.29	31.53	16.97	46.53	18.39	69.37	29.76
	Male	33.12	12.94	28.52	11.42	43.92	18.32	64.29	31.67
Group	Gender composition								
	All female	29.24	12.93	26.14	7.97	43.62	11.55	65.81	20.09
	Female majority	26.89	9.01	25.67	9.85	39.34	10.66	58.58	17.01
	Male majority	27.05	8.88	28.31	13.17	45.51	15.27	67.36	20.48
	All male	29.84	12.06	21.51	8.26	34.64	9.25	51.34	16.64

judgments and those from the mean, p's > 0.51, or the median, p's > 0.34, model. A comparison with the best-member model revealed significant differences for male-majority groups, $t(35) = 2.25$, $p = 0.03$, and female-majority groups, $t(34) = 2.61$, $p = 0.01$, but not for all-female groups, $t(34) = 0.66$, $p = 0.52$. Importantly, all-male groups were still significantly worse calibrated than the outcome of the mean, $t(35) = 5.96$, $p < 0.01$, the median, $t(35) = 5.66$, $p < 0.01$, or the best-member $t(35) = 8.20$, $p < 0.01$, model.

Judgment accuracy and confidence interval widths. Table 4 presents the absolute percentage errors as well as percentage interval widths aggregated over all 10 questions across the six decision-maker types for the 50%, 70%, and 90% confidence levels.

A one-way ANOVA of absolute percentage errors averaged over all 10 questions revealed a significant effect of decision-maker type, $F(5, 203) = 3.96$, $p < 0.01$. Planned contrasts did not show a significant difference in accuracy between all-male groups and the other groups, $F(1, 203) = 0.80$, $p = 0.37$. In contrast, a comparison of groups and individuals showed that group judgments were significantly more accurate than those of individuals, $F(1, 203) = 15.87$, $p < 0.01$, $d = 0.59$, but there was no significant difference between individual judgments made by men and women, $F(1, 203) = 2.48$, $p = 0.12$.

A 6 (decision-maker type) × 3 (confidence level) mixed ANOVA of percentage interval widths averaged over the 10 questions showed a significant main effect of both decision-maker type, $F(5, 406) = 3.75$, $p < 0.01$, and confidence level, $F(2, 406) = 575.80$, $p < 0.01$, but no significant interaction was observed, $F(10, 406) = 1.17$, $p = 0.31$. Joint tests across all three confidence levels with planned contrasts revealed that confidence intervals by all-male groups were significantly narrower than those of other groups, $F(3, 406) = 19.25$, $p < 0.01$, $d = 0.71$. Moreover, our analysis showed that there was no significant difference in interval widths between individual judgments by men and women, $F(3, 406) = 1.98$, $p = 0.12$. In addition, intervals provided by groups with at least one female member were significantly narrower than those provided by female individuals, $F(3, 406) = 4.77$, $p < 0.01$, $d = 0.31$, but not compared to those provided by male individuals, $F(3, 406) = 0.34$, $p = 0.80$. Intervals provided by all-male groups were significantly different from those stated by male, $F(3, 406) = 15.05$, $p < 0.01$, $d = 0.67$, or female, $F(3, 406) = 27.73$, $p < 0.01$, $d = 0.89$, individuals.

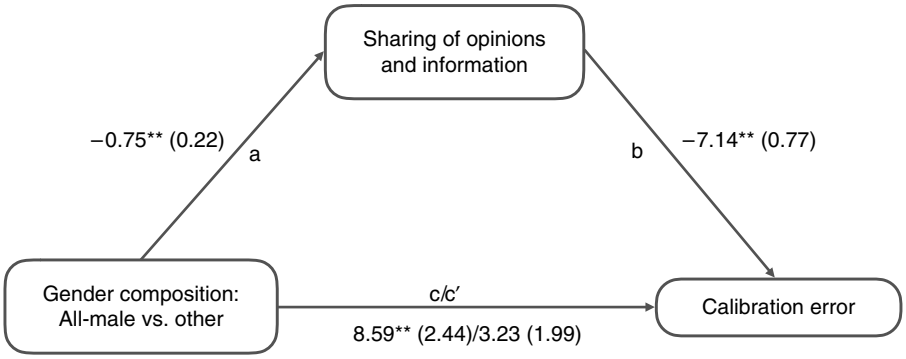
Mediation and group discussion. For each group member, we averaged the four items measuring the degree of opinion and information sharing into one composite measure ($\alpha = 0.88$) and then averaged the three group members' ratings into one aggregate measure for each group (ICC[1] = 0.55, $F[132, 266] = 4.67$, $p < 0.01$).[15] The results of a one-way ANOVA across group types revealed a significant main effect, $F(3, 129) = 4.30$, $p = 0.01$. Moreover, as predicted in Hypothesis 1, the results of a planned contrast indicate that members of all-male groups shared less information with each other than members of other groups, $F(1, 129) = 12.09$, $p < 0.01$, $d = 0.70$. Our results did not show a significant difference among other group types, p's > 0.42.

We next tested whether differences in calibration errors between all-male and other groups were mediated by the degree of opinion and information sharing. The results are shown in Figure 4.

Consistent with Hypothesis 1, our results revealed that the sharing of opinions and information was significantly lower in all-male groups (path a) than in other groups. Furthermore, the negative effect of having only male group members on calibration (path c) was reduced and became insignificant when we controlled for opinion and information sharing in the regression (path c and c′). Moreover, the 95% bootstrapped confidence interval for the mediated effect (2.27, 8.46) excluded zero, indicating significant mediation as predicted in Hypothesis 2B.

4.2.2. Results from Financial Forecasts. A comparison between return volatility estimates averaged

Figure 4. Results of Mediation Analysis (Study 2)



Notes. OLS regression coefficients. Standard errors in parentheses; $^{*}p < 0.05$; $^{**}p < 0.01$.

over all three time horizons and the mean historical return volatility (14.44%) revealed that both groups (M = 5.93%, SD = 3.06%), $t(141) = 33.10$, $p < 0.01$, and individuals (M = 6.62%, SD = 3.68%), $t(164) = 17.47$, $p < 0.01$, significantly underestimated return volatilities. The results of a 6 (decision-maker type) × 3 (time horizon) mixed ANOVA of return volatility estimates showed a significant main effect of time horizon, $F(2, 406) = 391.34$, $p < 0.01$, but not of decision-maker type, $F(5, 406) = 1.04$, $p = 0.40$, and there was no significant interaction between the two factors, $F(10, 406) = 0.47$, $p = 0.91$. However, planned contrasts conducted jointly across all three time horizons indicated that return volatility estimates by all-male groups were significantly lower (M = 5.14%, SD = 2.46%) than those by other group types (M = 6.21%, SD = 3.20%), $F(3, 406) = 3.98$, $p < 0.01$, $d = 0.35$. There was no significant difference between individual judgments by men (M = 6.59%, SD = 3.42%) and women (M = 6.64%, SD =3.95%), $F(3, 406) = 0.20$, $p = 0.90$. Similarly, there was no significant difference between groups with at least one female member and male, $F(3, 406) = 0.52$, $p = 0.67$, or female, $F(3, 406) = 1.05$, $p = 0.37$, individuals. In contrast, return volatility estimates by all-male groups were significantly lower than those by either male, $F(3, 406) = 4.81$, $p < 0.01$, $d = 0.49$, or female, $F(3, 406) = 5.30$, $p < 0.01$, $d = 0.46$, individuals.

4.2.3. Analysis of Audiotapes, Reported Satisfaction, and Perceived Interpersonal Sensitivity. Like in Study 1, in the following we focus on comparing all-male groups with other groups. A complete summary of all measures across the four group compositions is provided in Online Appendix Table A9. For all measures, we also tested for differences among the other three group types but found no systematic significant differences.

Group discussions lasted for 24 minutes on average (SD = 6.88). Although a one-way ANOVA did not indicate a significant difference in discussion length across different group types, $F(3, 138) = 2.01$, $p = 0.12$, a planned contrast showed that discussions in all-male groups (M = 22.82, SD = 5.65) took marginally significantly less time than those in other groups (M = 25.01, SD = 7.00), $F(1, 138) = 3.39$, $p = 0.07$, $d = 0.35$. Using the same procedure as described in Study 1, we next computed the variance of group member participation intensity that captures how evenly group members participated in the discussion. A one-way ANOVA revealed a significant difference in the participation variance across groups of different gender composition, $F(3, 138) = 3.10$, $p = 0.03$, and our follow-up analysis showed that the participation variance during the discussion was significantly larger in all-male groups (M = 0.036, SD = 0.028) than in other groups (M = 0.023, SD = 0.026), $F(1, 138) = 6.90$, $p = 0.01$, $d = 0.50$.

We next aggregated our three group member satisfaction items into one composite variable ($\alpha = 0.83$). A one-way ANOVA of average member satisfaction in each group revealed a significant difference across gender composition, $F(1, 129) = 3.80$, $p = 0.01$, and planned contrasts revealed that group members in all-male groups (M = 3.98, SD = 0.66) were significantly less satisfied than those in groups with at least one female member (M = 4.35, SD = 0.65), $F(1, 129) = 8.18$, $p = 0.01$, $d = 0.56$.

A one-way ANOVA across group types of the average group scores for our two measures of interpersonally (in)sensitive behavior, interruption and encouragement to participate, revealed a marginally significant main effect of group types for encouragement to participate, $F(3, 129) = 2.55$, $p = 0.06$, and interruption, $F(3, 129) = 2.40$, $p = 0.07$. Planned contrasts revealed that there were significantly more perceived interruptions in all-male groups (M = 3.26, SD = 0.71) than in other groups (M = 2.91, SD = 0.81), $F(1, 129) = 5.27$, $p = 0.02$, $d = 0.46$, and that there were significantly more perceived encouragements to speak in groups with at least one female member (M = 3.59, SD = 1.08) than in all-male groups (M = 3.13, SD = 0.83), $F(3, 129) = 5.27$, $p = 0.02$, $d = 0.45$.

We next analyzed our data from mixed-gender groups on an individual level with a 2 (rater gender) × 2

(target gender)×2 (group type: male-majority versus female-majority) mixed ANOVA.[16] We did not find significant differences in how male or female group members in mixed-gender groups were rated by other group members with respect to either interruptions, $F(1,390) = 0.30$, $p = 0.59$, or encouragements to speak, $F(1,390) = 0.00$, $p = 0.94$, suggesting that in mixed-gender groups, men and women showed similar levels of interpersonal sensitivity. We also did not find an effect of rater gender or any significant interaction effects, p's > 0.10.

Moreover, our results showed that men in mixed-gender groups were rated by their group members to be interrupting others less often, $F(1,402) = 6.14$, $p = 0.01$, $d = 0.25$, and encouraging others to speak more than men in all-male groups, $F(1,402) = 10.09$, $p < 0.01$, $d = 0.32$. Finally, our analysis showed that the level of perceived interpersonally sensitive behaviors by men in mixed-gender groups were similar to that of women in all-female groups for encouragements to speak, $F(1,396) = 0.37$, $p = 0.54$, and for interruptions, $F(1,396) = 0.57$, $p = 0.45$. Together, these findings indicate that, consistent with our theoretical framework, men display more interpersonally sensitive behavior in the presence of women than when they are interacting in all-male groups.

Summary. We found similar results as in Study 1 with respect to calibration errors and the mediating role of information sharing, providing overall supporting evidence to all our hypotheses. Contributing a number of additional insights, we also found that there were significantly more perceived interruptions and less perceived encouragements to speak in all-male groups than other group types. Moreover, male members in mixed-gender groups showed similar levels of interpersonal sensitivity as female group members, and higher levels of interpersonal sensitivity than men in all-male groups.

5. General Discussion

The results of two laboratory experiments revealed that confidence judgments by groups with at least one female member were significantly better calibrated than those by all-male groups. This effect was mediated by a higher degree of opinion and information sharing in groups with one or more female members. Consistently, our analysis of the audiotaped group discussion also established that in groups with at least one female member, group members participated more evenly in group discussions than those in all-male groups where discussions ended more quickly and were more likely to be dominated by a single member. In addition, our results in Study 2 revealed that group members in all-male groups were more likely to interrupt others and less likely to encourage others to participate than members in other groups. On the other hand, among groups with at least one female member, there was no difference in calibration, information sharing, or other group discussion measures.

In both studies, we did not find a significant difference in confidence calibration between judgments made by individual men and women, which is in line with some prior research (Biais et al. 2005, Jonsson and Allwood 2003) but different from others (Soll and Klayman 2004). Moreover, our results showed that confidence calibration in groups with at least one female member was generally at the same level or even better than what would be expected from a simple aggregation of individual judgments of the corresponding gender composition, but this did not hold for all-male groups whose calibration was actually worse than what would be expected from a statistical aggregation. Therefore, our results indicate that whereas group deliberation had either neutral or positive effects for groups with at least one female member, it was clearly detrimental for all-male groups. We suggest that this latter effect might be due to a lack of efficient sharing of divergent opinions and information during the group deliberation in all-male groups, which caused groups to set overly narrow confidence intervals. Due to this process, all-male groups might generally perform closer to the level of individual decision makers with respect to their confidence calibration than groups of other gender compositions where divergent opinions were more likely to be shared.

In general, our results from the mediation analysis, the statistical aggregation models, and the comparison of individual judgments by men and women all strongly indicate that, as we hypothesized, it is the group deliberation process, rather than group members' individual differences, that drives the difference in calibration between groups with at least one female member and all-male groups.

Consistent with the results from prior research (Plous 1995, Russo and Schoemaker 1992, Sniezek and Henry 1989), a direct comparison of group and individual judgments showed that groups with at least one female member on average tended to make significantly better calibrated judgments than individuals. However, this advantage of group decision making was mostly lost in the case of all-male groups. For men—when interacting in an all-male group—group discussions have a nonsignificant effect on calibration or even harm calibration compared to a simple aggregation of individual judgments.

Both accuracy and interval widths are factors that might be affecting calibration. Whereas we did not find a significant effect of groups' gender composition on judgment accuracy, confidence intervals set by all-male groups were significantly narrower than those set by other groups. Thus, whereas group gender composition did not significantly influence a group's ability

to correctly answer a particular question, our results indicate that it did have a strong effect on group members' appreciation of their own lack of knowledge. In particular, whereas all groups tend to be more accurate than individuals, this advantage is mostly offset in all-male groups by narrower confidence intervals, which is consistent with our theoretical predictions that compared with members in other groups, members in all-male groups are less aware of their limited knowledge because of a lack of information sharing.

Suggesting that our main findings are quite robust, we obtained consistent results with flat incentives in Study 1 and when participants' incentives were particularly linked to their answers in Study 2. Moreover, we found similar results with items that were simply adapted from prior work or were randomly sampled from two knowledge domains with which participants were relatively familiar.

Results from financial forecasts in both studies mirrored those from general-knowledge questions, providing converging evidence for our hypotheses from a different type of task. All-male groups provided confidence intervals that implied significantly lower return volatility in the stock market than those provided by groups with at least one female member. Moreover, whereas return volatility estimates by all groups were generally lower than those observed historically, volatility estimates by groups with only male members were even further away from historical volatilities. These findings also suggest that the detrimental effect resulting from the absence of female group members extends to tasks that are similar to those carried out within the finance industry—an area with a relatively high proportion of all-male groups.

In contrast to our results from the general-knowledge and financial forecast questions, in Study 1, we did not find a significant difference between all-male groups and other group types for estimates in the random-walk task, which provides an interesting boundary condition for the effects of group gender composition. An explanation for this outcome might be that, unlike the general-knowledge and financial forecast questions, this particular task requires mostly mathematical intuition and does not relate to real-world phenomena; therefore, even if group members strongly engage in the exchange of opinions and information, their lack of skill in a task might prevent them from taking advantage of this increase in available information (e.g., Woolley et al. 2010).

Our work makes several contributions. First, whereas prior research on overconfidence in groups (Plous 1995, Russo and Schoemaker 1992, Sniezek and Henry 1989) was limited to a direct comparison of individual and group judgments, we focus on the comparison between all-male groups and groups with at least one female member. In doing so, our study establishes gender composition as an important moderating factor that determines the extent to which group discussions can alleviate miscalibration in confidence judgments. In particular, our findings reveal that group deliberations have a neutral or positive effect on calibration for groups with one or more female members but actually harm calibration in all-male groups compared to a simple aggregation of individual judgments. In addition, our study also extends prior work on group confidence calibration to the area of financial forecasts. Our results from this domain suggest that our findings also have important practical implications. In particular, organizations in the financial sector that rely on such forecasts could attempt to improve the quality of their forecasts by adjusting their human resource practices to ensure that relatively small groups of analysts contain at least one female member.

Second, our study contributes to the literature on gender diversity in organizations. In particular, our findings demonstrate that the benefits arising from the presence of female group members could be more subtle than an increase in group performance—they might instead at least partially be driven by a lower susceptibility to judgmental biases such as overconfidence. In recent years, academic research, public media, and politics has paid considerable attention to the gender composition of top management teams and board of directors. Even though a higher share of women on boards and in top management teams is often considered desirable in the interest of gender equality, there have been mixed findings on its actual impact on firms' financial performance (e.g., Post and Byron 2015, Wolfers 2006). Our results suggest that one important advantage of avoiding all-male groups might be the increased ability to better deal with situations under substantial levels of uncertainty due to better confidence calibration. Such advantage might not be directly visible in firms' financial performance (which is also influenced by a large variety of other factors) but is crucial in keeping firms away from excessive risk, and hence away from the danger of bankruptcy (e.g., Ben-David et al. 2013). Moreover, whereas prior work has demonstrated the effect of gender diversity on boards and in top management teams on organizational outcomes such as financial performance (Post and Byron 2015), risk taking (Baixauli-Soler et al. 2015), and financial fraud (Cumming et al. 2015), little is known about the actual group processes that are driving these effects. By linking group gender composition to information sharing and the quality of group confidence judgments, our work provides insights into this "black box" and thus complements prior research in this area.

Third, our work adds further evidence to the extensive literature on the psychological processes triggered by group diversity (e.g., Mannix and Neale 2005,

Van Knippenberg and Schippers 2007). In particular, our findings concerning the beneficial effects of gender diversity on opinion and information sharing are in line with prior theoretical frameworks suggesting that the effects of group diversity do not predominantly derive from additional members' knowledge or skills, but rather from their impact on within-group processes such as information sharing and elaboration (e.g., Van Knippenberg and Schippers 2007, Van Knippenberg et al. 2004). It is, however, important to note that since our experimental evidence focuses specifically on the effects arising from the presence of women during group interactions and on the role of interpersonal sensitivity in changing the quality of group discussions, it is unclear the extent to which our findings might extend to other forms of diversity, such as demographic (e.g., ethnicity and age), functional, or educational diversity that have been extensively studied in the prior literature and have also been hypothesized to have similar effects on information processing (e.g., Van Knippenberg and Schippers 2007, Van Knippenberg et al. 2004). We also differ from this previous work in that we do not focus on exploring differences between diverse and homogenous groups, but instead mainly compare all-male groups and groups with at least one female member—including the case of homogenously female groups.

Related to this issue, previous literature in this area has also frequently pointed out that group diversity might in many cases lead to social categorization processes and fault lines within a group (e.g., Mannix and Neale 2005, Van Knippenberg and Schippers 2007), and could thus have a negative impact on performance and group member satisfaction. In contrast, our results showed little evidence for the presence of such a process that could potentially harm information sharing among group members. Moreover, we found that members of all-male groups actually displayed the lowest willingness to work with the other group members again. It would be an important topic for future research to identify the precise conditions under which gender diversity leads to problematic social categorization processes or lower group satisfaction.

Our findings also establish an interesting discontinuity in the effects of gender composition: whereas compared to groups with at least one female member, all-male groups were significantly worse calibrated and showed less opinion and information sharing during the group discussion, there was no significant difference in calibration or the extent of information sharing among all-female, female-majority, or male-majority groups. This is consistent with prior findings that female group members shape the nature of group discussions not only through their own behavior but also by affecting the behavior of male group members (e.g., Adams and Ferreira 2009, Williams and Polman 2015). Therefore, in a relatively small group as in our study, even the presence of just one woman in the group appears to be sufficient to derive all potential benefits.

Our work has several limitations. First, we focused only on small groups of three members; hence, our findings might not directly extend to larger groups. For example, according to prior work on minority status (e.g., Kanter 1977a, b), when minority members constitute less than 20% of the group, they are likely to be marginalized by other group members; thus, there might not be a positive effect on group outcomes arising from the presence of minorities. Moreover, social impact theory (Latane 1981) suggests that even in the absence of external marginalization from the majority group members, minority group members might still not be able to change the group dynamics if their number is too small compared to the total group size. These suggestions are also consistent with prior research on the effects of women on boards of directors, which has shown that even though the presence of only one woman might already be beneficial (e.g., Joecks et al. 2013, Zaichkowsky 2014), female board participation will only make its full contribution when the proportion of women reaches a certain "critical mass," which is often thought to be around 30%. In our study, except for all-male groups, women always comprised at least 33% of the entire group; thus, this threshold was always met. Given these prior findings, it would clearly be important to generalize our current findings to settings with larger groups and explore the influence of the required critical mass threshold in an experimental setting. Studying larger groups would also help to increase the external validity of our findings as managerial decisions are often made in groups with more than three members that, even independent from their gender composition, might exhibit different group norms than smaller groups.[17]

Another limitation of our study is that we only focused on a particular type of overconfidence. Future research should explore the effect of gender composition on other forms of overconfidence or other common cognitive biases, such as the escalation of commitment or the confirmation bias. Prior studies that have compared individuals and groups with respect to cognitive biases have reported very mixed results (e.g., Kerr et al. 1996). Our findings in this paper suggest that group gender composition might be an important moderating factor that could explain the circumstances under which groups deal with cognitive biases better than individuals. Similarly, it would be interesting to explore the effect of group gender composition on outcomes for of individual group members—for example, in the form of knowledge transfers from group discussions to subsequent individual judgments (e.g., Maciejovsky and Budescu 2007, Maciejovsky et al. 2013) or attitudes toward outgroups (e.g., Keck 2014).

Lastly, we cannot fully determine the exact processes driving the higher levels of opinion and information sharing in groups with at least one female member. In line with our theory, our findings indicate that members in groups with at least one female member showed higher levels of interpersonal sensitivity in group interactions such as fewer interruptions and more encouragements, both of which are associated with higher information sharing (Cooke and Szumal 1994, Leana 1985, Van Dyne and LePine 1998). However, there could theoretically still be factors other than the heightened interpersonal sensitivity that contribute to higher information sharing. First of all, as suggested by prior findings on sexual selection mechanisms (e.g., Griskevicius et al. 2006), it is possible that men in mixed-gender groups attempt to make themselves look more appealing to female members by talking more and thereby sharing more information. However, our findings—that group member participation is more balanced in groups with female members compared to all-male groups and that there is no difference in the participation variance among groups with zero, one, or two male members—stand in contrast to this suggestion.

Another possibility is that the higher group member satisfaction in groups with at least one female member might at least be a partial driver of the higher levels of participation and information sharing—instead of being a consequence of it. Although our findings do suggest that interpersonally sensitive behavior is at least partially driving higher information sharing, we are not able to fully determine the directional causal relationships between interpersonally sensitive behavior, group member satisfaction, and information sharing. In particular, it is possible that, for example, satisfaction and interpersonally sensitive behavior interact and reinforce each other over time and to some extent jointly drive information sharing. It would be interesting to test the precise relationship among these variables in an experiment by manipulating these factors independently from each other. This would provide more insights into this issue than our current work—which mainly focused on establishing the link between group gender composition, information sharing, and overconfidence, and less on the mechanisms that link the first two of these factors.

In conclusion, our findings indicate that, compared to all-male groups, the inclusion of female members significantly improves information sharing and, as a consequence, confidence calibration. Our findings have implications for research on group overconfidence and group judgments in general, as well as for other scholarly work on gender diversity in organizations and managerial practice. In particular, we highlighted mechanisms under which the inclusion of more women in top management teams or boards of directors might be beneficial for organizational outcomes.

Acknowledgments

The authors are grateful for constructive feedback from the three anonymous reviewers, the anonymous associate editor, and department editor Yuval Rottenstreich. They also thank David Budescu, Natalia Karelaia, Ilia Tsetlin, and members of the audience at the 2016 Academy of Management Annual Meeting for helpful comments. The authors also thank the Vienna Center for Experimental Economics for their support with conducting the two experiments.

Endnotes

[1] In both studies, we determined the sample size in advance and did not add or drop observations after we started our analyses other than where reported in the paper. We also did not collect data other than those reported in the paper.

[2] In neither experiment were any effects found arising from the gender of the person who entered the decisions.

[3] We also conducted our analysis using the signed differences and obtained overall very similar results.

[4] For most of our statistical analysis, we analyzed our data at the decision-maker level (either individuals or groups) by conducting mixed ANOVAs—using confidence levels (three levels) as a within-subject factor and decision-maker type (six types) as a between-subject factor. Here, degrees of freedom should be 148 (calculated as 154 – 6, where 154 refers to the total number of decision makers and 6 the number of decision-maker types) for the between-subject effect, and 296 (calculated as 148×2, where 2 refers to the number of confidence level minus one) for the within-subject effect and the interaction effect.

[5] Alternatively, in both studies we also conducted simple F-tests for which we averaged the results over all three confidence levels. This analysis gave consistent results, and our main findings remained significant.

[6] It is important to note that the average accuracy and interval width are only noisy predictors of the calibration error—e.g., a decision maker could be very inaccurate and have narrow intervals on one or two questions leading to an overall very high average percentage error and low interval width but for the rest of the questions provide intervals that contain the true values leading to an overall good calibration.

[7] In both studies, we also tested for differences between male and female members within mixed-gender groups with respect to all main measures discussed in the analysis but found no systematic significant differences.

[8] There were three observations for the Microsoft forecasts that resulted in negative volatilities. We removed these observations from the analysis. Including them does not change the significance of our main results.

[9] A detailed overview of our data for financial forecasts in both experiments and the random-walk task can be found in the online appendix.

[10] Although this difference was not significant, we found that intervals in all-female groups (M = 34.3, SD = 29.30) were considerably wider than in other groups. This might be because some members felt less confident about their understanding of the task because of gender stereotypes concerning mathematical abilities that might be more prominent in all-female groups.

[11] To determine the required sample size, we conducted a power analysis based on the observed effect sizes in Study 1. To achieve a satisfactory minimum power of 0.8 (e.g., Cohen 1992) with $\alpha = 0.05$ for a comparison of calibration errors between all-male groups and other group types averaged over all three confidence levels with a simple t-test, we would require a sample of size of about 36 in each

group. This would then also provide us with an excellent power of 0.97 to detect effects at the 90% confidence level where our effect was the strongest.

[12] We removed one observation from this condition as the participant's answers strongly indicated that she did not follow the experimental instructions.

[13] The choice of the two knowledge domains was guided by the result of a pretest with 68 students, which indicated that students considered themselves moderately knowledgeable with respect to both domains (M = 4.20 and M = 4.04 for distances and prices, respectively, on a 1 = "not at all" to 7 = "very much" scale) and that there were no significant differences in perceived knowledge between male and female students.

[14] Again, our choice of this item was guided by a pretest indicating at least a moderate level of familiarity (M = 3.70) and no significant gender differences.

[15] Due to an organizational mistake, we lost the data on opinion and information sharing for nine groups (three in the female-majority condition and two in each of the other conditions). Therefore, our following analysis is based on the remaining 133 groups.

[16] We also analyzed our two variables of interest with OLS regressions in which we clustered standard errors at the group level. This analysis gave results consistent with those provided here.

[17] Note, however, that even in boards and top management teams, decisions made by groups of three are not uncommon. For example, the size of the board of directors largely depends on the size of the firm, and smaller firms have been reported to have an average size of three to four (e.g., Bennedsen et al. 2008). Moreover, many board decisions are made by subcommittees that are only comprised of a small number of three to five board members. Similarly, strategic decisions are often made by a small subgroup of the top management team that consists of for example the CEO, COO, and head of a particular functional area, such as the CFO (e.g., Miles and Watkins 2007).

References

Adams RB, Ferreira D (2009) Women in the boardroom and their impact on governance and performance. *J. Financial Econom.* 94(2):291–309.

Anderson C, Brion S, Moore DA, Kennedy JA (2012) A status-enhancement account of overconfidence. *J. Personality Soc. Psych.* 103(4):718–735.

Anderson KJ, Leaper C (1998) Emotion talk between same-and mixed-gender friends—Form and function. *J. Language Soc. Psych.* 17(4):419–448.

Apesteguia J, Azmat G, Iriberri N (2012) The impact of gender composition on team performance and decision making: Evidence from the field. *Management Sci.* 58(1):78–93.

Asch SE (1952) Group forces in the modification and distortion of judgments. *Social Psychology* (Prentice Hall Englewood Cliffs, NJ), 450–501.

Baixauli-Soler JS, Belda-Ruiz M, Sanchez-Marin G (2015) Executive stock options, gender diversity in the top management team, and firm risk taking. *J. Bus. Res.* 68(2):451–463.

Barber BM, Odean T (2002) Online investors: Do the slow die first? *Rev. Financial Stud.* 15(2):455–488.

Baron RM, Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Personality Soc. Psych.* 51(6):1173–1182.

Bear JB, Woolley AW (2011) The role of gender in team collaboration and performance. *Interdisciplinary Sci. Rev.* 36(2):146–153.

Ben-David I, Graham JR, Harvey JR (2013) Managerial miscalibration. *Quart. J. Econom.* 122(3):1547–1584.

Bennedsen M, Kongsted HC, Nielsen KM (2008) The causal effect of board size in the performance of small and medium-sized firms. *J. Banking Finance* 32(6):1098–1109.

Berdahl JL, Anderson C (2005) Men, women, and leadership centralization in groups over time. *Group Dynam.* 9(1):45–57.

Biais B, Hilton D, Mazurier K, Pouget S (2005) Judgmental overconfidence, self-monitoring, and trading performance in an experimental financial market. *Rev. Econom. Stud.* 72(2):287–312.

Boschini A, Muren A, Persson M (2011) Men among men do not take norm enforcement seriously. *J. Socio-Econom.* 40(5):523–529.

Budescu DV, Du N (2007) Coherence and consistency of investors' probability judgments. *Management Sci.* 53(11):1731–1744.

Clemen RT (1996) *Making Hard Decisions: An Introduction to Decision Analysis* (PWS-Kent, Boston).

Cohen J (1992) A power primer. *Psych. Bull.* 112(1):155–159.

Cooke RA, Szumal JL (1994) The impact of group interaction styles on problem-solving effectiveness. *J. Appl. Behav. Sci.* 30(4): 415–437.

Cumming D, Leung T, Rui O (2015) Gender diversity and securities fraud. *Acad. Management J.* 58(5):1572–1593.

Davis-Stober CP, Budescu DV, Dana J, Broomell SB (2014) When is a crowd wise? *Decision* 1(2):79–101.

Deaves R, Luders E, Schröder M (2010) The dynamics of overconfidence: Evidence from stock market forecasters. *J. Econom. Behav. Organ.* 75(3):402–412.

Deutsch M (1949) An experimental study of the effects of cooperation and competition upon group process. *Human Relations* 2(3): 199–231.

Dezsö CL, Ross DG (2012) Does female representation in top management improve firm performance? A panel data investigation. *Strategic Management J.* 33(9):1072–1089.

Dufwenberg M, Muren A (2006) Generosity, anonymity, gender. *J. Econom. Behav. Organ.* 61(1):42–49.

Eagly AH, Johnson BT (1990) Gender and leadership style: A meta-analysis. *Psych. Bull.* 108(2):233–256.

Edmondson A (1999) Psychological safety and learning behavior in work teams. *Admin. Sci. Quart.* 44(2):350–383.

Fletcher J (1998) Relational practice: A feminist reconstruction of work. *J. Management Inquiry* 7(2):163–186.

Gaba A, Tsetlin I, Winkler RL (2017) Combining interval forecasts. *Decision Anal.* 14(1):1–20.

Gigerenzer G, Hoffrage U, Kleinbolting H (1991) Probabilistic mental models: A Brunswikian theory of confidence. *Psych. Rev.* 98(4):506–528.

Gigone D, Hastie R (1997) The impact of information on small group choice. *J. Personality Soc. Psych.* 72(1):132–140.

Glaser M, Langer T, Weber M (2013) True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *J. Behav. Decision Making* 26(5):405–417.

Gneezy U, Niederle M, Rustichini A (2003) Performance in competitive environments: Gender differences. *Quart. J. Econom.* 118(3):1049–1074.

Greenhalgh L, Chapman DI (1998) Negotiator relationships: Construct measurement, and demonstration of their impact on the process and outcomes of negotiation. *Group Decision Negotiation* 7(6):465–489.

Griskevicius V, Goldstein NJ, Mortensen CR, Cialdini RB, Kenrick DT (2006) Going along versus going alone: When fundamental motives facilitate strategic (non)conformity. *J. Personality Soc. Psych.* 91(2):281–294.

Hackman JR (1987) The design of work teams. Lorcsh J, ed. *Handbook of Organizational Behavior* (Prentice Hall Englewood Cliffs, NJ), 315–342.

Hall JA (1978) Gender effects in decoding nonverbal cues. *Psych. Bull.* 85(4):845–857.

Heilman ME (2012) Gender stereotypes and workplace bias. *Res. Organ. Behav.* 32:113–135.

Hinsz VB, Tindale RS, Vollrath DA (1997) The emerging conceptualization of groups as information processors. *Psych. Bull.* 121(1):43–64.

Hoogendoorn S, Oosterbeek H, Van Praag M (2013) The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Sci.* 59(7):1514–1528.

Hora SC (2004) Probability judgments for continuous quantities: Linear combinations and calibration. *Management Sci.* 50(5): 597–604.

Jain K, Mukherjee K, Bearden JN, Gaba A (2013) Unpacking the future: A nudge toward wider subjective confidence intervals. *Management Sci.* 59(9):1970–1987.

Janis IL (1982) *Groupthink: Psychological Studies of Policy Decisions and Fiascoes* (Wadsworth, Boston).

Jehn KA, Northcraft GB, Neale MA (1999) Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Admin. Sci. Quart.* 44(4):741–763.

Jehn KA, Rispens S, Thatcher SM (2010) The effects of conflict asymmetry on work group and individual outcomes. *Acad. Management J.* 53(3):596–616.

Joecks J, Pull K, Vetter K (2013) Gender diversity in the boardroom and firm performance: What exactly constitutes a critical mass? *J. Bus. Ethics* 118(1):61–72.

Jonsson AC, Allwood CM (2003) Stability and variability in the realism of confidence judgments over time, content domain, and gender. *Personality Individual Differences* 34(4):559–574.

Jose VRR, Winkler RL (2009) Evaluating quantile assessments. *Oper. Res.* 57(5):1287–1297.

Juslin P, Winman A, Olsson H (2000) Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psych. Rev.* 107(2):384–396.

Kanter RM (1977a) *Men and Women of the Corporation* (Basic Books, New York).

Kanter RM (1977b) Some effects of proportions on group life: Skewed sex ratios and responses to token women. *Am. J. Sociol.* 82(5):965–990.

Keck S (2014) Group reactions to dishonesty. *Organ. Behav. Human Decision Processes* 124(1):1–10.

Keefer DL, Bodily SE (1983) Three-point approximations for continuous random variables. *Management Sci.* 29(5):595–609.

Kennedy C (2003) Gender differences in committee decision-making: Process and outputs in an experimental setting. *J. Women Politics Policy* 25(3):27–45.

Kennedy JA, Anderson C, Moore DA (2013) When overconfidence is revealed to others: Testing the status-enhancement theory of overconfidence. *Organ. Behav. Human Decision Processes* 122(2): 266–279.

Kerr NL, MacCoun RJ, Kramer GP (1996) Bias in judgment: Comparing individuals and groups. *Psych. Rev.* 103(4):687–719.

Klayman J, Soll JB, Gonzalez-Vallejo C, Barlas S (1999) Overconfidence: It depends on how, what, and whom you ask. *Organ. Behav. Human Decision Processes* 79(3):216–247.

Kozlowski SW, Bell BS (2003) Work groups and teams in organizations. Borman WC, Ilgen DR, Klimoski RJ, eds., *Handbook of Psychology*, Vol. 12: Industrial and Organizational Psychology (Wiley, New York), 333–375.

Kozlowski SW, Klein KJ (2000) A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. Klein KJ, Kozlowski SW, eds. *Multilevel Theory, Research, and Methods in Organizations* (Jossey-Bass, San Francisco), 3–90.

Krishnan HA, Park D (2005) A few good women—On top management teams. *J. Bus. Res.* 58(12):1712–1720.

Latané B (1981) The psychology of social impact. *Amer. Psychologist* 36(4):343–356.

Laughlin PR, Ellis AL (1986) Demonstrability and social combination processes on mathematical intellective tasks. *J. Exp. Soc. Psych.* 22(3):177–189.

Laughlin PR, Bonner BL, Miner AG (2002) Groups perform better than the best individuals on letters-to-numbers problems. *Organ. Behav. Human Decision Processes* 88(2):605–620.

Leana CR (1985) A partial test of Janis' groupthink model: Effects of group cohesiveness and leader behavior on defective decision making. *J. Management* 11(1):5–18.

Levine JM, Smith E (2013) Group cognition: Collective information search and distribution. Carlston D, ed. *Oxford Handbook of Social Cognition* (Oxford University Press, New York), 616–633.

Lichtenstein S, Fischhoff B (1977) Do those who know more also know more about how much they know? *Organ. Behav. Human Perform.* 20(2):159–183.

Maciejovsky B, Budescu DV (2007) Collective induction without cooperation? Learning and knowledge transfer in cooperative groups and competitive auctions. *J. Personality Soc. Psych.* 92(5):854–870.

Maciejovsky B, Sutter M, Budescu DV, Bernau P (2013) Teams make you smarter: How exposure to teams improves individual decisions in probability and reasoning tasks. *Management Sci.* 59(6):1255–1270.

Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J. Personality Soc. Psych.* 107(2):276–299.

Mannix E, Neale MA (2005) What differences make a difference? The promise and reality of diverse teams in organizations. *Psych. Sci. Public Interest* 6(2):31–55.

Mast MS (2001) Gender differences and similarities in dominance hierarchies in same-gender groups based on speaking time. *Sex Roles* 44(9/10):537–556.

McLeod PL, Baron RS, Marti MW, Yoon K (1997) The eyes have it: Minority influence in face-to-face and computer-mediated group discussion. *J. Appl. Psych.* 82(5):706–718.

Miles SA, Watkins MD (2007) The leadership team. *Harvard Bus. Rev.* 85(4):90–98.

Minson JA, Mueller JS (2012) The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psych. Sci.* 23(3):219–224.

Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psych. Rev.* 115(2):502–517.

Nemeth CJ (1986) Differential contributions of majority and minority influence. *Psych. Rev.* 93(1):23–32.

Odean T (1998) Volume, volatility, price, and profit when all traders are above average. *J. Finance* 53(6):1887–1934.

Park S, Budescu DV (2015) Aggregating multiple probability intervals to improve calibration. *Judgment Decision Making* 10(2): 130–143.

Pearson ES, Tukey JW (1965) Approximate means and standard deviations based on distances between percentage points of frequency curves. *Biometrika* 52(3–4):533–546.

Phillips KW, Loyd DL (2006) When surface and deep-level diversity collide: The effects on dissenting group members. *Organ. Behav. Human Decision Processes* 99(2):143–160.

Plous S (1995) A comparison of strategies for reducing interval overconfidence in group judgments. *J. Appl. Psych.* 80(4):443–454.

Post C, Byron K (2015) Women on boards and firm financial performance: A meta-analysis. *Acad. Management J.* 58(5):1546–1571.

Russo JE, Schoemaker PJ (1992) Managing overconfidence. *Sloan Management Rev.* 33(2):7–17.

Schachter S (1959) *The Psychology of Affiliation* (Stanford University Press, Stanford, CA).

Shrout PE, Bolger N (2002) Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psych. Methods* 7(4):422–445.

Smith-Lovin L, Brody C (1989) Interruptions in group discussions: The effects of gender and group composition. *Amer. Sociol. Rev.* 54(3):424–435.

Sniezek JA (1990) A comparison of techniques for judgmental forecasting by groups with common information. *Group Organ. Management* 15(1):5–19.

Sniezek JA (1992) Groups under uncertainty: An examination of confidence in group decision making. *Organ. Behav. Human Decision Processes* 52(1):124–155.

Sniezek JA, Henry RA (1989) Accuracy and confidence in group judgment. *Organ. Behav. Human Decision Processes* 43(1):1–28.

Soll JB, Klayman J (2004) Overconfidence in interval estimates. *J. Experiment. Psych. Learning Memory Cognition* 30(2):299–314.

Stasser G (1992) Information salience and the discovery of hidden profiles by decision-making groups: A "thought experiment." *Organ. Behav. Human Decision Processes* 52(1):156–181.

Tindale RS, Larson JR (1992) Assembly bonus effect or typical group performance? A comment on Michaelsen, Watson, and Black (1989). *J. Appl. Psych.* 77(1):102–105.

Tost LP, Gino F, Larrick RP (2013) When power makes others speechless: The negative impact of leader power on team performance. *Acad. Management J.* 56(5):1465–1486.

Triandis HC, Kurowski LL, Gelfand MJ (1994) Workplace diversity. Triandis HC, Dunnette MD, Hough LM, eds. *Handbook of Industrial and Organizational Psychology*, Vol. 4 (Consulting Psychologists Press, Palo Alto, CA), 769–827.

Tsai CI, Klayman J, Hastie R (2008) Effects of amount of information on judgment accuracy and confidence. *Organ. Behav. Human Decision Processes* 107(2):97–105.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185:1124–1131.

Van Dyne L, LePine JA (1998) Helping and voice extra-role behaviors: Evidence of construct and predictive validity. *Acad. Management J.* 41(1):108–119.

Van Knippenberg D, Schippers MC (2007) Work group diversity. *Annual Rev. Psych.* 58:515–541.

Van Knippenberg D, De Dreu CK, Homan AC (2004) Work group diversity and group performance: An integrative model and research agenda. *J. Appl. Psych.* 89(6):1008–1022.

Van Vugt M, Iredale W (2013) Men behaving nicely: Public goods as peacock tails. *British J. Psych.* 104(1):3–13.

Wegge J, Roth C, Neubach B, Schmidt KH, Kanfer R (2008) Age and gender diversity as determinants of performance and health in a public organization: The role of task complexity and group size. *J. Appl. Psych.* 93(6):1301–1313.

Williams M, Polman E (2015) Is it me or her? How gender composition evokes interpersonally sensitive behavior on collaborative cross-boundary projects. *Organ. Sci.* 26(2):334–355.

Wolfers J (2006) Diagnosing discrimination: Stock returns and CEO gender. *J. Eur. Econom. Assoc.* 4(2):531–541.

Woolley AW, Chabris CF, Pentland A, Hashmi N, Malone TW (2010) Evidence for a collective intelligence factor in the performance of human groups. *Science* 330(6004):686–688.

Zaichkowsky JL (2014) Women in the board room: One can make a difference. *Internat. J. Bus. Governance Ethics* 9(1):91–113.

Zarnoth P, Sniezek JA (1997) The social influence of confidence in group decision making. *J. Experiment. Social Psych.* 33(4): 345–366.