Demographically diverse crowds are typically not much wiser than homogeneous crowds

Stephanie de Oliveira[a,1] and Richard E. Nisbett[a,1]

[a]Department of Psychology, University of Michigan, Ann Arbor, MI 48109

Contributed by Richard E. Nisbett, December 27, 2017 (sent for review October 16, 2017; reviewed by Hal R. Arkes and Jack B. Soll)

Averaging independent numerical judgments can be more accurate than the average individual judgment. This "wisdom of crowds" effect has been shown with large, diverse samples, but the layperson wishing to take advantage of this may only have access to the opinions of a small, more demographically homogeneous "convenience sample." How wise are homogeneous crowds relative to diverse crowds? In simulations and survey studies, we demonstrate three necessary conditions under which small socially diverse crowds can outperform socially homogeneous crowds: Social identity must predict judgment, the effect of social identity on judgment must be at least moderate in size, and the average estimates of the social groups in question must "bracket" the truth being judged. Seven survey studies suggest that these conditions are rarely met in real judgment tasks. Comparisons between the performances of diverse and homogeneous crowds further confirm that social diversity can make crowds wiser but typically by a very small margin.

judgment accuracy | demographic diversity | estimate aggregation

The quality of people's decisions depends in part on the quality of their judgments. To improve judgment, popular intuition recommends considering multiple people's opinions, and research supports this recommendation (1). One particularly impressive phenomenon, the "wisdom of crowds," demonstrates the effectiveness of combining multiple opinions for numerical judgments about matters of fact. Averages of individual estimates (i.e., "crowds") can outperform the average individual in that crowd. The accuracy gains are most dramatic for small crowds; averaging one's own estimate with just two other people's estimates can decrease error by over 30% (2).

This "wisdom" is typically observed when estimates are independent and randomly chosen to be aggregated by some method like averaging. This often allows people's errors to cancel out during the averaging process, as each person's guess is comprised of truth plus some positive or negative error. For example, suppose two people are predicting the high temperature in Tampa tomorrow, and the correct prediction is 83. Person A may guess 85, so he has a positive error as he overestimates the true value. Person B may guess 80, so she has a negative error as she underestimates the true value. When their guesses are averaged, the errors cancel out to some degree since they bracket the truth, and that crowd is wiser than the average individual in that crowd. Specifically, the crowd guess (82.5) is off by 0.5°. Person A is off by 2° and person B is off by 3°, so the average individual is off by 2.5°. When bracketing is not observed—everyone overestimates or underestimates the truth—then the average person is as wise as the crowd (3).

Social influence can significantly weaken the wisdom of crowds effect because as people interact with each other's guesses they converge in their thinking, making their error more systematic (that is, they become biased in the same way) (4). If people make their estimates more independently, this preserves diversity in their errors, which supports the canceling that leads to greater accuracy.

Effective wisdom of crowds has typically been demonstrated by researchers who can randomly sample from a diverse pool of people to create crowds. When wisdom of crowds is observed, it signifies that those people had diverse errors. How would nonscientists fare in creating wise crowds given that they are unlikely to randomly ask people for advice? Wisdom of crowds can be observed with just a few people (1, 5), so crowd size is not a resource issue. However, might there be a homophily problem? For example, could a crowd help a white, college-educated female if it is comprised of her office mates—other white, college-educated females? Or should she make an effort to ask other colleagues who differ in gender, ethnicity, or educational attainment?

The idea that cognitive diversity improves crowd judgment is well supported (6, 7). Cognitive diversity—variation in people's judgments or how they think—is hard to directly assess. Thus, people commonly use social diversity as a proxy for cognitive diversity, expecting people who differ externally to also differ cognitively (8, 9). These differences are commonly expected to translate into performance benefits, as reflected in these statements from popular news outlets: "…diversity on boards improves decision-making and profits" (*The New York Times*, ref. 10). "Need a balance sheet boost? Try adding some women to the board of directors" (*Los Angeles Times*, ref. 11).

Diversity scholars take a more nuanced and cautious approach than laypeople when discussing the benefits of diversity (12–15). Nevertheless, it is not unreasonable to believe that social category membership influences how one thinks across a variety of topics. Race influences perceptions of social interactions (16), culture influences predictions of stock market trends (17), one's favored sports team influences one's predictions about game outcomes (18), and political values influence interpretations of the facts in a court case (19). More broadly, for cognitive (2) or

Significance

Leveraging social diversity to maximize group performance is an important challenge in the organizational sphere. Reported studies examine the effects of demographic diversity on the accuracy of "crowd" judgment—statistically aggregated individual judgments. Results suggest that demographic diversity does not boost crowds' cognitive diversity to the extent necessary to make diverse crowds much wiser than homogeneous ones. A strong implication is that a decision to seek diverse opinion on matters of fact should be based on a cost/benefit analysis: Will a search for diversity likely pay off in increased accuracy? Payoffs can be maximized by using stronger correlates of cognitive diversity than demographic variables.

Author contributions: S.d.O. and R.E.N. designed research; S.d.O. performed research; S.d.O. analyzed data; and S.d.O. and R.E.N. wrote the paper.

Reviewers: H.R.A., Ohio State University; and J.B.S., Fuqua School of Business.

The authors declare no conflict of interest.

Data deposition: Data from the survey studies are available in the Deep Blue Data repository (https://doi.org/10.7302/Z2D50K5H).

[1]To whom correspondence may be addressed. Email: sdeochen@umich.edu or nisbett@umich.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717632115/-/DCSupplemental.

Published online February 9, 2018.

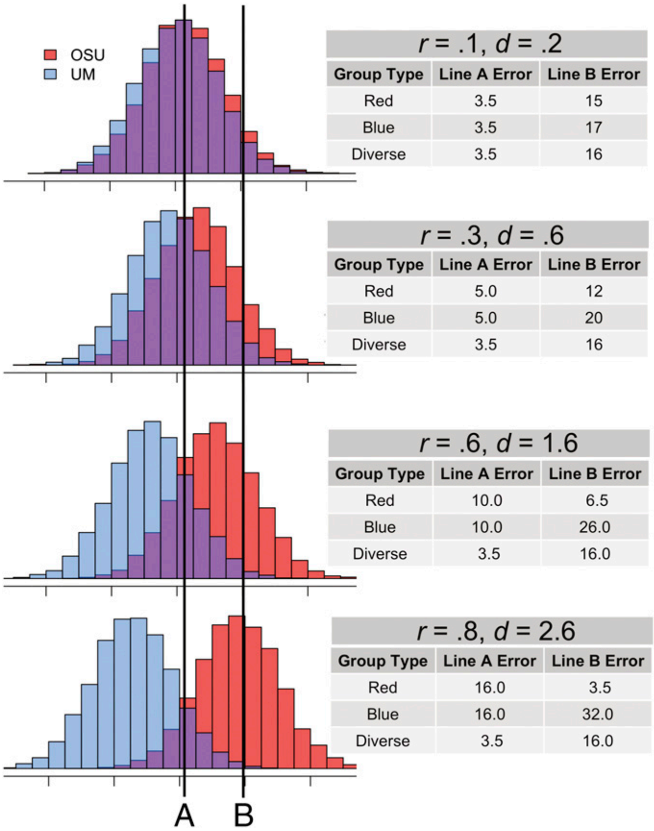
motivational reasons (20), people make judgments anchored around their own views, experiences, and identities (21, 22). But for numerical judgments with a correct answer, is the social–cognitive connection warranted? Or does the assumption that it would amount to inaccurate stereotyping (23)? In other words, to what extent do people overestimate the link between a social attribute (e.g., being a conservative, being female) and judgment about matters of fact (e.g., making biased predictions in favor of one's preferred candidate or overestimating the popularity of a "chick flick")?

In Theory, Social Diversity Can Boost Crowd Wisdom

In theory, social diversity can make crowds wiser. Three conditions must all be met. First, the social–cognitive connection must be real; the social characteristic of concern must systematically influence judgments. For example, men and women must reliably make different estimates for a given type of problem. If sex differences do not map onto cognitive differences, then it would make no difference whether one uses aggregates of mixed sex or only one sex.

Second, the group difference must be large enough to have a meaningful impact on the truth bracketing rate and, by extension, accuracy. At small effect sizes (e.g., r = 0.1), the distribution of men's and women's guesses would overlap substantially, even if the groups' mean differences were statistically significant. The groups would be so similar that a small sample including men and women would hardly differ from a small sample of only men or only women. Fig. 1 illustrates this principle by showing hypothetical estimates from two groups that overlap to varying degrees.

Third, the correct answer must lie between the average estimate of each group. In other words, the distribution of estimates from each social group must somewhat evenly bracket the truth (3). This allows for each group to have the same level of (in)accuracy but in opposite biased directions relative to the truth. Their biases can then neatly cancel out in a diverse group. (Imagine if, instead, the distribution of women's guesses is relatively close to the truth and the distribution of men's guesses is relatively farther away. A diverse crowd would not produce the wisest crowd; one would be better off just asking women for their estimates in that case.)



r = .1, d = .2

Group Type	Line A Error	Line B Error
Red	3.5	15
Blue	3.5	17
Diverse	3.5	16

r = .3, d = .6

Group Type	Line A Error	Line B Error
Red	5.0	12
Blue	5.0	20
Diverse	3.5	16

r = .6, d = 1.6

Group Type	Line A Error	Line B Error
Red	10.0	6.5
Blue	10.0	26.0
Diverse	3.5	16.0

r = .8, d = 2.6

Group Type	Line A Error	Line B Error
Red	16.0	3.5
Blue	16.0	32.0
Diverse	3.5	16.0

Fig. 1. Error of homogeneous and diverse crowds. From *Top* to *Bottom*, the distributions represent small to large effects of social identity on judgment. Purple regions reflect overlap of red and blue groups. Line A represents an outcome that meets the bracketing condition; line B represents an outcome that violates the bracketing condition.

The Model: Diverse Crowd Accuracy as a Function of Effect Size and Bracketing

We modeled the wisdom of homogeneous and diverse crowds from two simulated, hypothetical social groups to demonstrate the above requirements. In the model, the effect size of "social diversity" on judgment was manipulated, as was the location of the truth relative to the groups' guesses. As these parameters varied, the relative accuracy of diverse (vs. homogeneous) crowds varied. The manipulations, which we refer to as "effect size" and "bracketing," are described below.

Effect Size. We simulated crowds by sampling from one or two normal distributions (Fig. 1). Each distribution represented the population estimates from a homogeneous social group. As an illustration, imagine one distribution represents estimates from Ohio State (OSU) fans and one distribution represents estimates from University of Michigan (UM) fans. Imagine that they are forecasting how many points OSU will score in an upcoming game. In the model, creating a homogeneous social crowd means drawing estimates from only one of the distributions (e.g., asking only UM fans for their estimates). Creating a diverse social crowd means drawing estimates from both distributions—asking four OSU fans and four UM fans for their estimates.

Note that how much the groups' estimates overlap predicts how much one stands to gain by employing a diverse crowd versus a homogeneous crowd. If the effect of social group membership on estimates is small (r = 0.1), the OSU and UM distributions overlap a lot, almost to the point of being the same. Diverse crowd performance will on average be nearly indistinguishable from homogeneous crowd performance. If the effect of group membership on estimates is large (r = 0.8), the OSU, UM, and diverse crowd distributions will be far more distinct.

Four pairs of distributions were created to represent the following effect sizes: r = 0.8, d = 2.6 (M_{OSU} = 58 and M_{UM} = 26); r = 0.6, d = 1.6 (M_{OSU} = 52 and M_{UM} = 32); r = 0.3, d = 0.6 (M_{OSU} = 46 and M_{UM} = 38); and r = 0.1, d = 0.2 (M_{OSU} = 43 and M_{UM} = 41). All SDs were 12.25, and the midpoint between each distribution's mean was always 42. (The numbers are arbitrary; other values can be used to create distributions at each effect size. We chose 42 because that is how many points OSU really scored in the 2015 match against UM, and 12.25 was roughly the standard deviation in a study we ran on football forecasts.) Each distribution contained 10,000 "estimates" from that social group. To simulate crowds, we averaged eight guesses from homogeneous vs. diverse sampling as described above. Twenty thousand homogeneous crowds were created (10,000 OSU crowds and 10,000 UM crowds), and 10,000 diverse crowds were created.

Bracketing. How well each crowd performs for a given question also depends on where the true value lies. Diverse crowds will be wiser than homogeneous crowds when the outcome lies near the midpoint between OSU's and UM's respective mean guesses (Fig. 1, line A). The farther away the truth is from the midpoint (e.g., Fig. 1, line B), the less of an advantage diversity offers. In the model, accuracy is the absolute difference between the crowd's estimate and the outcome (i.e., absolute error). Each crowd's performance was compared against varied hypothetical outcome values. Accuracy was first computed with the truth (T)

set to the midpoint, 42 (Fig. 1, line A). It was then recomputed at T = 43, 44, 45, and so forth until T = 62. The full results are graphed in Figs. S1 and S2, where the x axis represents movement of the truth away from the midpoint between M_{OSU} and M_{UM}, the y axis represents absolute error, and the lines represent homogeneous and diverse crowds. Summarized results for $T = 42$ (line A) and $T = 58$ (line B) are shown in the tables included in Fig. 1. The errors are rounded to the nearest 0.5, but the exact observed errors from the simulation are plotted in Figs. S1 and S2.

The errors reported in Fig. 1 suggest that diverse crowd superiority over homogeneous crowds is a function of both effect size and bracketing. Consider results for line A, where the bracketing condition is met. At small effect sizes ($r = 0.1$, Cohen's $d = 0.2$), diverse crowds of eight are not more accurate than homogeneous crowds. When $r = 0.3$ ($d = 0.6$), diverse crowds begin to exhibit an advantage over homogeneous ones. When $r = 0.6$ and 0.8 ($d = 1.6$ and 2.6), diverse crowds have a more substantial advantage. But when bracketing is violated (line B), the diversity advantage is lost for r = 0.3, only to reemerge slightly at $r = 0.6$ and more robustly at $r = 0.8$.

Another notable result is that when bracketing is violated (line B), the wisest crowd is a homogeneous one. If one can identify the wisest homogeneous crowd with some level of reliability, then one would be better off utilizing a homogeneous crowd over a diverse one (24). But if one has no basis for identifying the wisest homogeneous crowd, one should compare the diverse crowd's error to the average homogeneous crowd error (i.e., expected error). Further analyses reporting how accurate one must be at identifying the better homogeneous crowd are reported in *Probability of Identifying Wiser Crowd*.

In summary, choosing diverse crowds is a relatively effective strategy at larger effect sizes, as (*i*) the bracketing condition is harder to violate (there is more room for the truth to move and still be closer to the midpoint than to the mean of either group) and (*ii*) the estimates of the different groups are more distinct. But at smaller effect sizes, estimates from different groups are not very distinct and the bracketing condition is easier to violate given the narrower range at which the truth may lie.

In Practice, Diversity Minimally Boosts Crowd Accuracy

The simulations show that social diversity can make crowds wiser when the social factor is at least moderately associated with judgment and bracketing occurs. How often are those conditions met? Large effects of social identity or values have been found for some judgment tasks, but these typically ask about polarized attitudes (e.g., ref. 25) or questions for which the truth is subjective or cannot be known with certainty (e.g., ref. 19). Such effect sizes may not represent the typical effect of social identity on judgments about matters of fact. They may also not represent the typical effect size in social psychology more generally.

A meta-analysis spanning 100 years of psychological research suggested that the "overall average" effect size in social psychology is $r = 0.21$ (26). The most prevalent social factor the authors discussed was gender. Some effects of gender on social behavior and attitudes were moderate; women gaze at others more (0.29) and are more empathetic (0.37) than men. They are more likely to support the feminist movement (0.39). But most gender effects were much smaller. For example, gender differences in social attribution were all smaller than 0.08. Female jurors are slightly harsher than male jurors in sexual assault cases (0.16). Men are slightly more likely than women to dislike gays (0.04–0.19). Boys are more competitive than girls by a small margin (0.03), and men have higher self-esteem than women (0.06). Finally, race effects on judgment were also small: Whites report higher life satisfaction than African Americans (0.10). It is notable that of the few judgment topics reported, none were about matters of fact.

We are unaware of any research that systematically tested for demographic differences in numerical judgments about matters of fact as opposed to attitudes and subjective beliefs. It is reasonable to anticipate that such effects would be even smaller than the effects of social identity on attitude judgments and social behavior, because giving numerical answers to factual questions imposes a type of accountability on the respondent. One might say, "Men are a lot taller than women," but when pressed to define what "a lot" means, one is faced with the possibility of objective assessment of one's accuracy. People's biases are often reduced when there is accountability for one's answers (20, 27, 28).

Effect Sizes Are Small. Across nine judgment tasks spanning seven studies, we rarely observed moderate to large effects of social identity on judgment, even when making a serious effort to match judgment questions with social identity. In the studies, we measured social factors and correlated them against judgments in a wide array of domains. Social diversity can refer to myriad categories—gender, education, affinity for bird watching. Studies were mostly limited to demographic categories because these are frequently discussed in university and company diversity statements and, in one review, account for almost 90% of studied diversity effects (29). Specifically, we mostly measured age, sex, ethnicity or cultural background, educational attainment, religion, and political orientation. We also examined fans from opposing teams in one study. Correlations were computed between social factors and judgments in different domains. Football fans from opposing teams guessed how many points each team would score in a rivalry game in study 1. In studies 2 and 4, respectively, people of diverse political backgrounds predicted how presidential candidates would perform in primary and national elections. In study 3, people of diverse political backgrounds guessed the level of national support for six different political statements. They also gave likelihood ratings for presidential candidates winning the upcoming Iowa caucus. In study 5, participants guessed the popularity ratings of 24 different novels. In studies 6 and 7, people forecasted the probability of outcomes for up to 40 diverse news stories in the United States and abroad (e.g., "What is the likelihood of Chinese official GDP growth exceeding 6.3% in the first quarter?" "What is the likelihood of Leonardo DiCaprio winning an Oscar?"). In study 7, they also predicted medal awards for the 2016 Olympics and forecasted future stock prices. The surveys were done in accordance with the University of Michigan Institutional Review Board guidelines, and surveys were completed after informed consent.

Hardly any strong relationships emerged when we explored all possible judgments against all measured social variables. Out of 965 correlations, 60% were weaker than $r = \pm 0.1$ and 96% were weaker than $r = \pm 0.2$. Fewer than 1% of the correlations were 0.3 or stronger. These magnitudes suggest that for most of these questions it would be impossible for socially diverse crowds to substantially outperform homogeneous crowds. All correlations significant at the $P < 0.01$ level are reported in Table S3.

Diverse Crowds Resemble Homogeneous Crowds. We compared the accuracy of socially homogeneous and diverse crowds from people's answers in the above studies. To give diverse groups the best chance of outperforming homogeneous groups, we only analyzed people's guesses for the tasks that showed the largest effects of social identity on judgment. They included sports predictions, political questions, and guesses about popular interest in various books. The social identity factors analyzed were likewise diverse, including the sports teams people cheered for, age, sex, political orientation, and whether they were religious.

From people's estimates, 1,000 crowds were created by averaging eight randomly selected people's estimates. Diverse crowds included four people from each social category (e.g., four liberals and four conservatives). Likewise, homogeneous crowds were created for each social category (e.g., 1,000 crowds including eight randomly chosen liberals and 1,000 crowds including eight randomly chosen conservatives). Individual and crowd accuracy was assessed by comparing mean absolute errors (MAEs). People

Table 1. Accuracy of individuals, socially homogeneous crowds, and diverse crowds on judgment tasks

A	B	C	D	E	F	G	H	I	J	K
Task	Study N	Social identity	r	Average individual error (SD)	Homogeneous crowd 1 error	Homogeneous crowd 2 error	Average homogeneous crowd error	Diverse crowd error	Average homogeneous crowd error if perfect bracketing	Average diverse crowd error if perfect bracketing
1	51	Team	0.36	14.49 (5.25)	9.88_{OSU}	17.32_{UM}	13.6	13.52	3.98	2.44
2	198	Religiosity	0.13	14.46 (4.65)	12.46_{REL}	11.24_{NREL}	11.85	11.79	3.62	3.38
3	614	Age	0.11	18.19 (6.05)	9.9_{YOUNG}	10.22_{OLD}	10.06	9.9	6.21	5.81
4	234	Political orientation	0.21	14.52 (6.70)	9.35_{LIB}	8.47_{CON}	8.91	7.99	5.67	4.44
5	205	Age	0.1	2.24 (0.57)	1.09_{YOUNG}	1.26_{OLD}	1.18	1.14	0.92	0.88
5	205	Sex	0.1	2.24 (0.57)	1.05_{MEN}	1.11_{WOM}	1.07	1.06	0.86	0.83

A: The tasks were as follows: (*i*) predict points in OSU vs. UM game, (*ii*) predict percentage of votes eight presidential candidates would receive in two state primaries, (*iii*) guess what percentage of Americans support each of six political statements, (*iv*) predict what percentage of votes Clinton and Trump would each win in 10 states in the 2016 presidential election, and (*iv*) guess the popularity rating that 24 diverse books received in a previous study. C: Social identity is the social factor that was most strongly associated with people's answers. D: The *r* is the average (absolute) correlation between that social category and answers. E: Average individual error is the average of absolute error across all questions in the task. F–I: These are absolute errors for different types of crowds. F and G are homogeneous crowds, H is the average between homogeneous crowds, and I is diverse crowds. J and K: J is what homogeneous crowd error would have been if perfect bracketing had been observed, and K is the same for diverse crowds.

often do not know a priori which, if any, homogeneous crowd will be the most accurate, so the simulated diverse crowds are compared against the average homogeneous crowd's performance.

Comparing columns E, H, and I in Table 1 suggests that both homogeneous and diverse crowds performed much better than the average individual across studies. But diverse crowds often performed about as well as the average homogeneous crowd. The largest effect of social diversity on judgment was for sports: People's favored sports team had a moderate effect on their predictions for an upcoming game. However, the bracketing condition was not met; the game's outcome surprised everyone: OSU did better than most people expected, and UM did worse. [These participants only made two target judgments—1 per team. A better test of the bracketing condition would involve multiple judgments as in studies 2 through 7. We therefore asked 110 fans to guess the points scored across 20 past games for a total of 40 judgments. However, the effect of favored team on judgment disappeared, not significantly differing from 0. If the large effect were observed across many judgments, it is likely that the bracketing condition would be met at least some of the time, affording diverse crowds some advantage over homogeneous ones.] This resulted in diverse crowds performing about as well as homogeneous crowds. For task 4—predicting presidential candidate performance—social diversity decreased error the most. The socially diverse crowd, comprised of liberals and conservatives, was typically 1% point closer to the true outcome than the average homogeneous crowd comprised of only liberals or only conservatives. This amounted to a 10% decrease in error, from 8.91 to 7.99.

The small effect sizes mean that the margin by which diverse crowds could possibly outperform homogeneous crowds is low, even with perfect bracketing. But bracketing was typically violated by several points on the response scale. Columns J and K report what homogeneous and diverse crowd error would have been if, for every question, the outcome or true answer had been the midpoint between the mean estimate of each homogeneous group. In that highly unlikely best-case scenario, diverse crowds outperform homogeneous ones for the tasks with the largest effect sizes (tasks 1 and 4) but offer minimal advantages for the other tasks.

"Very Homogeneous" vs. "Very Diverse" Crowds. A stronger test of the diversity hypothesis would be to test the accuracy of very homogeneous groups against very diverse groups. For example, a group of religious, white Republicans is more homogeneous than a group of religious people because its members overlap on multiple social categories. We examined such crowds' performance for the same tasks in Table 1 except for the football task, as the sample size was too small.

For these analyses, very homogeneous pools of participants were created by including people who overlapped on two or three social dimensions. Six homogeneous pools were created a priori for each study based on factors that could reasonably be argued as relevant to the judgment (Table 2). For example, during American election cycles, news media frequently refer to age, sex, ethnicity, social class, and political orientation as factors that divide people into distinct voting blocks. Thus, homogeneous pools were created around those criteria. The pools were restricted by sample size; only those for which at least 30 participants met the criteria were used (with two exceptions, Table 2 legend). To create a diverse pool from which to simulate crowds, 40 participants were randomly chosen from the complete dataset of each study except in study 3. Given the larger *N*, we were able to create larger homogeneous pools, so the diverse pool included 100 participants.

Homogeneous and diverse aggregates of eight people were created 1,000 times. Accuracy, or MAE, of the two types of aggregates was compared, with the homogeneous groups averaged together. Performance broken down by crowd type can be seen in Fig. 2, where G1 always represents diverse crowds and G2–G7 represent homogeneous crowds. For primary election judgments, $MAE_{G1} = 11.56$ and $MAE_{G2–7} = 11.96$. For popular political opinion judgments, $MAE_{G1} = 9.40$ and $MAE_{G2–7} = 9.61$. For presidential election forecasts, $MAE_{G1} = 9.09$ and $MAE_{G2–7} = 8.89$. For book popularity judgments, *MAE* of G1 and G2–7 were both 1.12. Overall, diverse crowds did not consistently perform much better than the average homogeneous crowd. (Note that we expect that if all types of homogeneous crowds could be created and sampled against diverse crowds, the error of homogeneous crowds could not be lower than the error of diverse crowds as it was for the presidential election forecasts).

Fig. 2 shows some slight variation in homogeneous group performance. Thus, choosing a diverse crowd can be considered a risk reduction strategy, even if it is unlikely to be the best-performing group. A diverse crowd is essentially guaranteed a moderately good performance, whereas choosing a homogeneous crowd introduces the possibility of doing relatively well and also the possibility of

Table 2. Very homogeneous and diverse groups

Task	G1*	G2	G3	G4	G5	G6	G7
Predict percentage of votes eight presidential candidates would receive in two state primaries	Random	White men, did not complete college	White women, completed college	Religious white Republican	Nonreligious white Democrats	Liberal women under 40	Liberal nonwhites
Guess what percentage of Americans support each of six political statements	Random	White men, did not complete college	White women, completed college	Religious white conservative	Nonreligious white liberals	Liberal women under 40	Liberal nonwhites
Predict what percentage of votes Clinton and Trump would each win in 10 states in 2016 presidential election	Random	White men, did not complete college	White women, completed college	Religious white Republican	Nonreligious white Democrats	Liberal women over 40	Liberal nonwhites
Guess the popularity rating that 24 diverse books received in a previous study	Random	Men over 40	Men under 30	Women over 40	Women under 30	Ethnic minority women	White men

G1* is always the diverse crowd. All groups except two were simulated from pools of at least 30 people. G2 for the book task (men over 40) was simulated from a pool of 22 men, and G6 (ethnic minority women) was sampled from a pool of 29 due to limited representation of those groups in the larger sample.

doing relatively poorly. If one has no basis for choosing one homogeneous crowd over another, a diverse crowd is a "safe" choice.

Discussion

Social diversity can theoretically make crowds wiser but only to the extent to which it corresponds to diversity in estimates. People's estimates must not only be diverse, they must also bracket the true value in at least some realizations of the judgment outcome. We replicate previous findings demonstrating that small crowds can be wiser than individuals but find that it matters far less whether the crowd is homogeneous or diverse along a given demographic dimension.

These results suggest several conclusions. First, for numerical judgments, socially homogeneous crowds may produce about as much estimate diversity as socially diverse crowds. Given the small effect sizes observed, the diversity between groups was relatively small and the diversity within groups was relatively large. Second, people who expect social groups to think very differently for these types of judgments may be erroneously stereotyping, expecting variation within groups to be small and variation between groups to be large. Indeed, people tend to expect external, social factors to signal internal, cognitive attributes (8, 9). Our work supports exhortations to avoid stereotyping errors in diversity discussions (20). This might be achieved by emphasizing that for numerical judgments about matters of fact, there is a lot of cognitive diversity within homogeneous social groups (30); not all women think alike, not all liberals think alike, and so forth.

Practically, when choosing between diverse or homogeneous crowds for utilitarian reasons, one must consider (*i*) the likelihood and magnitude of accuracy gains over individual judgment, (*ii*) how valuable gains of those magnitudes are for the judgment at hand, and (*iii*) the relative costs of obtaining diverse versus homogeneous estimates. This paper sheds light on *i*. We conclude that accuracy gains with diverse crowds are unlikely to be much larger than gains with homogeneous crowds because we only observed diversity gains for a minority of the tasks we studied. However, decision makers must factor *ii* and *iii* into their deliberations. They should also consider other reasons for pursuing diversity that are based on equity and representation. Our findings are consistent with meta-analyses and reviews suggesting no straightforward performance benefit for using diverse versus homogeneous groups (31–33), and they bolster the argument that people often overstate the performance case for diversity (15). But pursuing demographic diversity on boards, juries, and other influential decision-making teams helps ensure that the interests and values of a diverse population are fairly represented and addressed.

A second practical consideration relates to generalizability. People in teams do a lot more than make numerical judgments. Judgments are a key but small part of larger, more complex tasks like decision-making and problem-solving, and team members often interact in noncrowdsourcing contexts. Therefore, caution is warranted when generalizing the present findings to other contexts where the task, team coordination, and performance metrics are different. As we note in the previous paragraph, the present findings are consistent with null diversity benefits reported in the literature (31–33). But some work suggests that encountering demographic diversity can improve cognition at the individual level (34–36). For example, white participants scored higher on reading comprehension and memory measures when they read

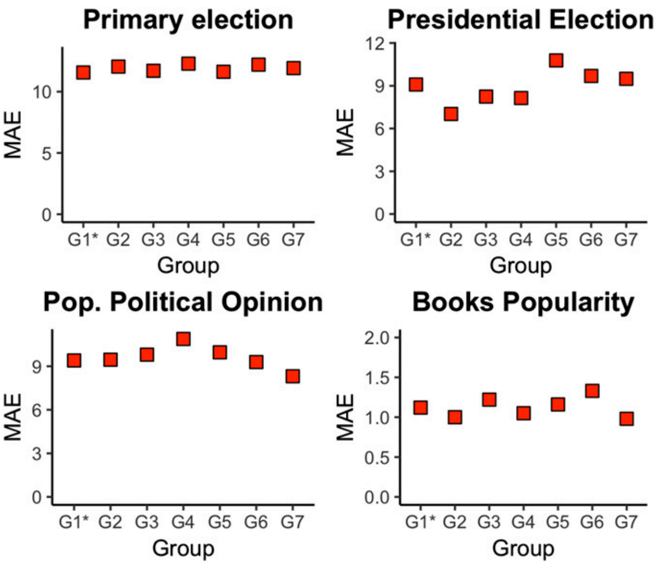


Fig. 2. Very homogeneous vs. very diverse groups across four tasks. G1 always represents the diverse crowds. G2 to G7 represent the homogeneous groups as described in Table 2. For example, for the primary election, presidential election, and popular political opinion tasks, G2 refers to white men who did not complete college. For the book-rating task, G2 refers to men over 40 y old. In all graphs, the *y* axis indicates error. Lower values mean higher accuracy on the task.

study materials in racially diverse discussion groups as opposed to all-white groups (36). Researchers should further explore whether and how those individual cognitive improvements can translate into improved performance at the group level.

Our findings do not mean that no social factors correlate with any types of judgment; in many cases, social identity can influence more subjective judgment. In addition, there is good evidence that culture affects many aspects of cognition from perception to problem solving (37). Our results do suggest that according to the operationalizations of social diversity in the present work, which are consistent with many popular conceptions of "diversity," there is typically only a weak correlation between demographics and numerical judgment. Those correlations are insufficient for making diverse crowds much more accurate than homogeneous ones. For crowds, we would expect that any measure that effectively taps into people's systematic biases would be a better proxy for cognitive diversity than the demographic variables that are commonly used.

ACKNOWLEDGMENTS. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant DGE 1256260, awarded to S.d.O.

1. Yaniv I (2004) The benefit of additional opinions. *Curr Dir Psychol Sci* 13:75–78.
2. Yaniv I, Milyavsky M (2007) Using advice from multiple sources to revise and improve judgments. *Organ Behav Hum Decis Process* 103:104–120.
3. Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Manage Sci* 52:111–127.
4. Lorenz J, Rauhut H, Schweitzer F, Helbing D (2011) How social influence can undermine the wisdom of crowd effect. *Proc Natl Acad Sci USA* 108:9020–9025.
5. Mannes AE, Soll JB, Larrick RP (2014) The wisdom of select crowds. *J Pers Soc Psychol* 107:276–299.
6. Page SE (2008) *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies* (Princeton Univ Press, Princeton).
7. Davis-Stober CP, Budescu DV, Broomell SB, Dana J (2015) The composition of optimally wise crowds. *Decis Anal* 12:130–143.
8. Northcraft GB, Polzer JT, Neale MA, Kramer RM (1995) Diversity, social identity, and performance: Emergent social dynamics in cross-functional teams. *Diversity in Work Teams: Research Paradigms for a Changing Workplace*, eds Jackson SE, Ruderman MN (American Psychological Association, Washington, DC), pp 69–96.
9. Phillips KW, Loyd DL (2006) When surface and deep-level diversity collide: The effects on dissenting group members. *Organ Behav Hum Decis Process* 99:143–160.
10. Miller CC (June 19, 2014) Women on boards: Quotas have limited success. *The New York Times*. Available at www.nytimes.com/2014/06/20/upshot/women-on-the-board-quotas-have-limited-success.html?abt=0002&abg=1. Accessed November 1, 2017.
11. Hsu T (August 1, 2012) Women on board: Firms with female directors do better, study says. *Los Angeles Times*. Available at articles.latimes.com/2012/aug/01/business/la-fi-mo-women-board-performance-20120801. Accessed November 1, 2017.
12. Mannix E, Neale MA (2005) What differences make a difference? The promise and reality of diverse teams in organizations. *Psychol Sci Public Interest* 6:31–55.
13. Jehn KA, Northcraft GB, Neale MA (1999) Why differences make a difference: A field study of diversity, conflict, and performance in workgroups. *Adm Sci Q* 44:741–763.
14. Klein KJ, Harrison DA (2007) On the diversity of diversity: Tidy logic, messier realities. *Acad Manag Perspect* 21:26–33.
15. Eagly AH (2016) When passionate advocates meet research on diversity, does the honest broker stand a chance? *J Soc Issues* 72:199–222.
16. Dovidio JF, Kawakami K, Gaertner SL (2002) Implicit and explicit prejudice and interracial interaction. *J Pers Soc Psychol* 82:62–68.
17. Ji L-J, Zhang Z, Guo T (2008) To buy or to sell: Cultural differences in stock market decisions based on price trends. *J Behav Decis Mak* 21:399–413.
18. Simmons JP, Massey C (2012) Is optimism real? *J Exp Psychol Gen* 141:630–634.
19. Kahan DM (2010) Culture, cognition, and consent: Who perceives what, and why, in 'acquaintance rape' cases. *Univ PA Law Rev* 158:729–813.
20. Kunda Z (1990) The case for motivated reasoning. *Psychol Bull* 108:480–498.
21. Ross L, Greene D, House P (1977) The 'false consensus effect': An egocentric bias in social perception and attribution processes. *J Exp Soc Psychol* 13:279–301.
22. Krueger J, Clement RW (1994) The truly false consensus effect: An ineradicable and egocentric bias in social perception. *J Pers Soc Psychol* 67:596–610.
23. Page SE (2007) Making the difference: Applying a logic of diversity. *Acad Manag Perspect* 21:6–20.
24. Soll JB, Larrick RP (2009) Strategies for revising judgment: How (and how well) people use others' opinions. *J Exp Psychol Learn Mem Cogn* 35:780–805.
25. Graham J, Haidt J, Nosek BA (2009) Liberals and conservatives rely on different sets of moral foundations. *J Pers Soc Psychol* 96:1029–1046.
26. Richard FD, Bond CF, Stokes-Zoota JJ (2003) One hundred years of social psychology quantitatively described. *Rev Gen Psychol* 7:331–363.
27. Windschitl PD, Smith AR, Rose JP, Krizan Z (2010) The desirability bias in predictions: Going optimistic without leaving realism. *Organ Behav Hum Decis Process* 111:33–47.
28. Lerner JS, Tetlock PE (1999) Accounting for the effects of accountability. *Psychol Bull* 125:255–275.
29. Jackson SE, Joshi A, Erhardt NL (2003) Recent research on team and organizational diversity: SWOT analysis and implications. *J Manage* 29:801–830.
30. Park B, Hastie R (1987) Perception of variability in category development: Instance-versus abstraction-based stereotypes. *J Pers Soc Psychol* 53:621–635.
31. van Dijk H, van Engen ML, van Knippenberg D (2012) Defying conventional wisdom: A meta-analytical examination of the differences between demographic and job-related diversity relationships with performance. *Organ Behav Hum Decis Process* 119:38–53.
32. Webber SS, Donahue LM (2001) Impact of highly and less job-related diversity on work group cohesion and performance: A meta-analysis. *J Manage* 27:141–162.
33. van Knippenberg D, Schippers MC (2007) Work group diversity. *Annu Rev Psychol* 58: 515–541.
34. Levine SS, et al. (2014) Ethnic diversity deflates price bubbles. *Proc Natl Acad Sci USA* 111:18524–18529.
35. Loyd DL, Wang CS, Phillips KW, Lount RB (2013) Social category diversity promotes premeeting elaboration: The role of relationship focus. *Organ Sci* 24:757–772.
36. Sommers SR, Warp LS, Mahoney CC (2008) Cognitive effects of racial diversity: White individuals' information processing in heterogeneous groups. *J Exp Soc Psychol* 44: 1129–1136.
37. de Oliveira S, Nisbett RE (2017) Culture changes how we think about thinking: From "Human Inference" to "Geography of Thought". *Pers Psychol Sci* 12:787–790.
38. Mellers B, et al. (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psychol Sci* 25:1106–1115.
39. Tetlock PE, Gardner D (2015) *Superforecasting: The Art and Science of Prediction* (Crown, New York).