The Perceived and Actual Influence of Social Diversity on “Crowd” Judgment Accuracy

by

Stephanie de Oliveira Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Psychology)
in the University of Michigan
2017

Doctoral Committee:

Professor Richard E. Nisbett, chair
Professor David A. Dunning
Professor Jeffrey Sanchez-Burks
Professor J. Frank Yates

Stephanie d. Chen

sdeochen@umich.edu

ORCID iD: 0000-0002-8615-2819

ACKNOWLEDGEMENTS

I could not have completed the work in this dissertation without the help and encouragement of many. I would like to first thank Richard Nisbett for adopting me when I was adviser-less. Things could not have turned out better for me. I would not be remotely as good a scholar or as wise a person without your excellent mentorship. I am also enormously grateful for Frank Yates' insightful advising and friendship, and for Phoebe Ellsworth's abundant good sense while discussing just about any research idea.

Other colleagues have also been consistently brilliant and helpful. Thank you Decision Lab and PR lab folks, especially Haoyang Yan, Neil Lewis, and Mike Hall for your insights. Thank you Josh Wondra for offering statistics help with patience and clarity. Thank you also to Shinobu Kitayama and the Culture and Cognition lab for supporting my early work on culture. I am also thankful for Rick Larrick's expert guidance on the studies in Chapter IV, and for David Dunning's perceptive critiques of Chapters II and III.

Many wonderful people helped me navigate life these past five years. Thank you Doyle and Crystal, Pat Chen, Yay-hyung Cho, "Stephanie C. Senior", Jackie Kim, and Emily Wilson for great conversations and memories. I'm also immensely grateful for Christ Church's care for my family and the support of Northwood families.

Finally, I am so grateful to my family. Kevin, you have been there from the beginning. Thank you for pushing me to apply to Michigan, for your unwavering love and support, and for frequent comedic relief. You help me ask important questions and keep my perspective realistic.

I couldn’t have done this without you. Mom and dad, thank you for loving me, nurturing my curiosity, and teaching me to work hard. Miranda, thanks for caring for Isaque with so much love! Isaque, thank you for brightening every day with your smile. You are my sunshine.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF TABLES v

LIST OF FIGURES vi

LIST OF APPENDICES vii

ABSTRACT viii

CHAPTER

I. Introduction 1

II. Conditions for Diversity Accuracy Gains 15

Modeling Social Diversity and Cognitive Diversity 15

Results 19

III. Observed Diversity Accuracy Gains 22

Studies III.1 – III.7 23

IV. Lay Expectations of Diversity Gains 34

Study IV.1 35

Study IV.2 46

Study IV.3 53

Study IV.4 56

Study IV.5 62

V. General Discussion 70

APPENDICES 76

BIBLIOGRAPHY 86

LIST OF TABLES

TABLE

III.1 – Demographic Information, Studies III.1-III.7 23

III.2 – Study Summaries 24

III.3 – Accuracy of Individuals, Socially Homogeneous Crowds, and Diverse Crowds on Judgment Tasks 27

III.4 – Socially Homogeneous Groups 29

IV.1 – Summary of Studies IV.1a – IV.1c 35

IV.2 – People’s Estimate of Other People’s Predictions 41

B.1 – Correlations Between Social Identity and Judgment, $p < .01$. 84

LIST OF FIGURES

FIGURE

II.1 – Four Effect Sizes for Social Identity and Judgment 17

II.2 – Accuracy as a Function of Crowd Type, Effect Size, and Bracketing 19

II.1 – Error of “Very Diverse” and “Very Homogeneous” Crowds 30

IV.1 – Answer Format, Study 1c 39

IV.2 – Bias by Bracketing (2 X 2) Design 47

IV.3 – Effect of Bias and Bracketing on Diverse Choice 49

IV.4 – Distribution of Maximum WTP by Advice Type, Study IV.4 60

IV.5 – Distribution of Maximum WTP by Advice Type, Study IV.5 66

LIST OF APPENDICES

APPENDIX

A – Chapter III Studies: Methods Details 76

B – Chapter III Significant Correlations 83

ABSTRACT

Statistically aggregating even a few estimates can yield improvements over individual judgments. But before aggregating, one must first decide whom to ask. One strategy might be to ask a socially diverse group, as diverse people are expected to contribute different perspectives. Related to this idea, the aims of this dissertation were threefold. The first aim was to determine the necessary conditions under which socially diverse crowds outperform homogeneous crowds for a given numerical judgment. The second aim was to test to what extent these conditions are met in real judgment tasks across a variety of social identities and judgment questions. The third aim was to determine how realistic laypeople are when appraising the accuracy advantages – or lack thereof – of socially diverse crowds.

Chapter I reviews relevant literature on judgment, groups, and diversity. The discussion makes clear that the question of whether social diversity boosts crowd accuracy cannot be answered by the extensive literature on group diversity. The chapter nevertheless reviews that work and uses it to guide hypotheses while at the same time drawing distinctions between the questions asked here and the questions previously answered.

A model in Chapter II tests when socially diverse crowds will outperform homogeneous ones. Findings suggest that diversity only improves crowd judgment when the relationship between group membership and judgment is at least moderate in size and when the true value lies between the distributions of the two groups. Chapter III then seeks to observe these conditions in real judgment tasks. Results indicate that the conditions for diversity benefits are rarely observed in empirical data. People’s social identities did not strongly bias their judgments across a wide variety of topics, and homogeneous and diverse crowds performed about equally well on numerical judgment tasks.

Studies IV.1-IV.5 examined lay beliefs about diversity advantages. People typically overestimated the relative accuracy of socially diverse groups over homogeneous groups. That overestimation arose from people’s erroneous assumptions about meeting the “diversity benefit conditions” in the model (Studies IV.1 – IV.2). Specifically, people (1) overestimated the effect of social identity on judgment and (2) expected bracketing to occur to some degree. People expected diverse crowds to be optimal when they assumed these conditions are present, but not when they assumed that those conditions were absent.

Studies IV.3-IV.5 suggest that people are willing to act on those erroneous beliefs. They are far more likely to choose advice from a diverse crowd, even when a homogeneous crowd would be more accurate (Study IV.3), and they are willing to pay more money to see advice from a diverse crowd than a homogeneous one when completing a numerical judgment task (Study IV.4-IV.5). Finally, Chapter V reviews the findings in the dissertation, discusses implications and limitations of the present work, and proposes future studies to address remaining questions.

CHAPTER I

Introduction

Improving judgment accuracy is a major step towards improving decision quality. To improve accuracy, popular intuition recommends considering multiple people's opinions: "Two heads are better than one." Implementing this strategy has never been easier thanks to advances in communication and collaboration technologies. But optimal implementation remains a challenge; how can someone make a group as "wise" as possible? Simply convening a group of experts is one obvious possibility, but experts are not always readily identifiable or available, especially to the layperson.

Another popular strategy for maximizing the effectiveness of "many heads" is to include "diverse heads" (Galinsky et al., 2015; Page, 2008). There are many different kinds of diversity, but people often mean social, demographic diversity when they discuss its utility for groups and teams (McGrath, Berdahl, & Arrow, 1995; Northcraft, Polzer, Neale, & Kramer, 1995; Page, 2008). The present work therefore specifically tests how beneficial social diversity is for improving group performance above and beyond the performance of socially homogeneous groups. I focus on one particular task – numerical judgment – and the "groups" refer to statistical combination of people's estimates after people have individually made their judgments. The present work also explores Americans' lay theories about social diversity benefits in numeric judgment tasks.

Group composition and coordination

Assuming that knowledge and skills are differentially distributed among people, a group is likely to have more skills and abilities at its disposal than a given individual. Whether a group performs at its full potential depends on a variety of factors, such as what kind of diversity is present in the group, how the group members coordinate their contributions, the nature of the task at hand, and what is meant by "performance." In the present work, "performance" refers to numerical judgment accuracy, and the nature of the task is numerical judgment as opposed to other activities like problem solving. The next sections describe what I mean by "diversity" and "coordination."

Group diversity. Group composition can be described in many ways, such as according to its members' specializations. Research suggests that this type of diversity – cognitive diversity – is beneficial for group performance (Page, 2008). The present work focuses on social diversity, broadly conceived, because it is often what people mean when they refer to "diversity" (McGrath et al., 1995). Thus socially "diverse" groups include people that are different along a social, demographic factor like sex, cultural background, or educational attainment. Homogeneous group members all share a social identity (e.g., all are men).

Social diversity is often promoted on the grounds that it will enhance group performance by boosting cognitive diversity in a group. People and organizations often make this social-cognitive diversity connection (Northcraft et al., 1995; Page, 2007), but is it warranted? Research supports the connection to some degree. Race influences perceptions of social interactions (Dovidio, Kawakami, & Gaertner, 2002), culture influences predictions of stock market trends (Ji, Zhang, & Guo, 2008), one's favored sports team influences one's sports predictions (Simmons & Massey, 2012), and political factors influence interpretations of the facts in a court case (Kahan, 2010). More broadly, for cognitive (e.g., Yaniv & Milyavsky, 2007) or motivational reasons (e.g., Kahan, 2016; Kunda, 1990), people make judgments anchored around their own views, experiences, and identities (Krueger & Clement, 1994; Ross, Greene, & House, 1977). Nevertheless, scholars are generally cautious about touting social diversity's benefits, emphasizing the complexity of social diversity and its influence on group performance (Eagly, 2016; Jehn, Northcraft, & Neale, 1999; Klein & Harrison, 2007; Mannix & Neale, 2005; Page, 2007).

The present work empirically tests connections between social diversity and cognitive diversity – specifically, numerical judgments – and then tests social diversity's impact on group accuracy. Distinguishing between social and cognitive diversity is important; if social diversity does not significantly correspond to cognitive diversity, then people of different groups will be making judgments that are redundant with each other, producing no increase in accuracy.

Group coordination. In this paper, by "group" I mean the mathematical aggregation of individual judgments. The groups refer to collections of people who do not interact as they make their judgments, but instead each provide their own best judgment independently. Then, those judgments are aggregated (e.g., the mean is calculated) in order to produce the group's judgment. I will refer to these mathematically created, nominal groups as "crowds".

There are several reasons one might prefer crowds over using real, deliberating groups. First, deliberating groups often fail to reach their full potential (Sunstein & Hastie, 2015). In tasks where the information critical to making an accurate judgment is dispersed among different group members, people tend to spend more time discussing shared information rather than offering up unshared, critical information (Stasser & Stewart, 1992). This, in turn, leads to poor group judgment. Moreover, deliberating groups have been shown to render judgments that are the mere aggregates of pre-deliberation views (Gigone & Hastie, 1993) – in other words, deliberating groups acted as if they were statistical groups, making deliberation a waste of time. Even worse, other studies suggest that deliberation increases confidence in judgment without increasing accuracy (Heath & Gonzalez, 1995).

Various factors can contribute to group failures (for a thorough discussion, see Sunstein & Hastie, 2008; Sunstein & Hastie, 2015). Social factors can hinder information sharing and critical thinking. For example, dissenting group members may feel too intimidated to share information that is incongruent with the majority view, or a group member may assume that others already know what she knows (Asch, 1951; Kaplan & Miller, 1987). Value diversity can increase conflict and decrease morale (Jehn, Northcraft, & Neale, 1999). Cognitive factors can play a role, too. For example, people may misremember the critical unique information that they share with the group, significantly hampering group accuracy (Lightle, Kagel, & Arkes, 2009). Additionally, the order in which information is shared in a group can lead to incorrect conclusions based purely on the available information rather than on social pressures (Hung & Plott, 2001; Kaplan & Miller, 1987). When working independently, group members can avoid these social and cognitive pitfalls that hinder group performance.

A second reason for using crowds is that they are cheap. When team members all work on a problem individually, they each spend more time on task than if they had to convene together and discuss a host of matters like how to go about the task and who will be in charge. Crowd members can work wherever they wish, so the time and energy costs associated with corralling members into a single meeting are avoided.

A third reason for using such crowds is that they often perform very well relative to individuals; several heads usually *are* better than one. When certain statistical conditions are met, aggregation of a group's estimates can markedly outperform the average individual group member (Larrick & Soll, 2006; Mannes, Soll, & Larrick, 2014). In fact, the less exposure group members have to each others' estimates, the more accurate the aggregate estimate is likely to be (Lorenz, Rauhut, Schweitzer, & Helbing, 2011).

How "crowds" work

The phenomenon in which a group aggregate outperforms average individuals has been called the "wisdom of the crowds" effect (Surowiecki, 2005), hence the use of the term "crowds" in this paper. This effect arises from bias cancelation processes that occur during aggregation. In numerical judgment, each estimate can be thought of as the sum of three elements: truth, bias (systematic error associated with an individual), and random error. Aggregates created by averaging are most accurate when their errors are diverse, leading to some overestimation and some underestimation in the distribution – that is, when their constituent values "bracket" the true value (Davis-Stober, Budescu, Broomell, & Dana, 2015; Herzog & Hertwig, 2014; Larrick & Soll, 2006).

For example, Person A may guess that tomorrow's high temperature will be around 80 degrees whereas her colleague, B, may guess it will be around 70. If the true temperature lies between both estimates at 75 degrees, the accuracy of the aggregate (taken by averaging) will be perfect, although A and B will individually each be off by 5 degrees. With opposing biases leading to bracketing, the average estimate outperforms the average individual in the group. Had both guesses been biased in the same direction – say, 80 and 78 degrees – the average would not perform as well. In fact, the aggregate's error (4 degrees) would be as great as the expected error of randomly choosing A's estimate or B's colleague's estimate; A's error would have been 5 degrees, and B's would have been 3, the average of which is 4.

Social diversity and crowds

Estimate diversity is good for crowds, but how can it be consistently achieved? Researchers who study wisdom of crowds effects typically randomly sample individuals to create aggregates. These random samples include people who are demographically and cognitively diverse. Laypeople who wish to harness wisdom of crowds may only have access to a few people who are similar to them – family members, coworkers, or friends. Might that social similarity correspond to estimate similarity, creating a homophily problem? Should a layperson seek out socially diverse “crowds”? Or does cognitive/estimate diversity operate independently of social diversity?

There is a large literature on diversity and group performance. Much of that work uses different definitions of diversity, groups, and performance than those used in this paper. Consequently, that work does not directly answer the question presented here. However, I review that literature here to summarize general findings, justify the types of social category diversity used in the present work, and to clearly show the distinction between the present work and what has been done before.

Types of diversity. General definitions of diversity in the organizational literature are often similar: “the distribution of personal attributes among interdependent members of a work unit” (Jackson et al., 2003, p. 802), or “the distribution of differences among the members of a unit with respect to a common attribute, *X*, such as tenure, ethnicity, conscientiousness, task attitude, or pay” (Harrison & Klein, 2007, p. 1200). Some researchers also discuss the configuration of the distribution of those differences as an important feature of diversity (Harrison & Klein, 2007; Lau & Murnighan, 1998). More concrete operationalizations of diversity vary by study, however. One review found that most prior work focused on immediately apparent demographic features like sex, ethnicity, and age as opposed to deeper, less visible types of diversity like informational diversity (Jackson, Joshi, & Erhardt, 2003). Diversity in those three demographic categories accounted for 89% of reported diversity effects. Ironically, demographic diversity is frequently believed by theorists to be less task-relevant than other types of diversity (Pelled, 1996; Pelled, Eisenhardt, & Xin, 1999). Only 24% of reported effects examined more task-relevant diversity such as educational diversity.

Scholars frequently divide operationalizations of diversity into sub-types. In addition to the distinction between less task-relevant versus more task-relevant diversity (see also Webber & Donahue, 2001), others have similarly divided diversity into "surface level" diversity – that which is immediately apparent – versus "deep level" diversity, that which indicates different opinions, information, and values (e.g. Phillips & Loyd, 2006). Yet others have examined informational diversity (what people know), social category diversity (e.g., demographic variables), and value diversity (differences in goals and preferences) (Jehn et al., 1999).

The distinction in the literature that has the most bearing on the present work is between social or demographic diversity (who people are on the outside) and cognitive diversity (who people are inside their heads) (Page, 2008). This is similar to the "surface level" and "deep level" distinction above. Both are examined in the present work, as is further described below.

Types of performance. The definition of "performance" also varies between studies and disciplines. When studies are conducted on employees in organizational settings, the performance measurements can include financial performance (Jackson et al., 2003), evaluations from managers and supervisors (Jehn et al., 1999; Pelled et al., 1999), and metrics common in the particular industry. For example, one study used NACDA Director's Cup Points to evaluate NCAA Division I-A athletic departments' performance (Cunningham, 2009). Experimental studies in psychology often use psychological and social performance measures such as integrative complexity (Antonio et al., 2004), the quality of solutions in problem solving tasks (Homan, van Knippenberg, Van Kleef, & De Dreu, 2007), reading comprehension (Sommers, Warp, & Mahoney, 2008), or quality of group discussions (Phillips & Loyd, 2006; Phillips, Mannix, Neale, & Gruenfeld, 2004; Sommers, 2006).

Diversity's effects on performance. Reviews typically conclude that sometimes diversity helps task performance, sometimes it hurts, and sometimes it has no effect (Harrison & Klein, 2007; Jackson et al., 2003; van Knippenberg & Schippers, 2007). This partly arises from variation in definitions of diversity and performance. For example, Jehn and colleagues (1999) examined the effects of social category diversity, value diversity, and informational diversity on team performance across an organization. Performance was provided by the company's departmental records and included measures of team members' ratings of each other, team efficiency, and some objective productivity measures. They found that informational diversity was positively associated with performance, but social category diversity had no association with performance. Value diversity had a negative relationship with performance. In short, diversity's effects on performance depended on what kind of diversity was being measured.

In discussing organizational diversity, Webber and Donahue (2001) observed that social diversity is usually described as a "double-edged" sword, supposedly making groups perform *better* on tasks while performing *worse* on measures of group cohesion. They tested the validity of that narrative in a meta-analysis that analyzed the effects of highly work-related diversity and less work-related diversity on cohesion and performance. They hypothesized, based on the common narrative, that task-related diversity such as diversity in people's functional backgrounds should improve task performance, while less task-related diversity such as demographic diversity should have no influence on performance. Moreover, demographic diversity should decrease cohesion while having minimal effects on actual performance. To their surprise, *neither* type of diversity predicted either outcome measure.

Newer studies suggest, however, that social diversity can help groups when organizations have strategies in place to leverage it. Cunningham (2009) examined whether racial diversity predicted athletic department performance in the NCAA as measured by NACDA Director's Cup points. Results suggested that simply having a racially diverse department did not boost the department's performance. Moreover, having a diversity management strategy did not in and of itself boost performance. (Their measure of "strategy" included factors like taking a broad view of diversity, valuing diversity, and having open lines of communication.) However, having *both* diversity and a management strategy was associated with higher performance. Thus, diversity was beneficial if there were measures in place to capitalize on its strengths.

Some research also suggests that diversity helps when people think it will help, although the effect appears to be linked to informational diversity, not demographic diversity. In a study by Homan and colleagues (2007), gender-diverse teams completed a task. Teams espousing pro-diversity beliefs did not outperform teams that held pro-similarity beliefs. But the authors also manipulated how information was distributed such that for some teams it was diversely distributed – men had different information than women – and for others it was evenly distributed across members regardless of gender. For teams that had diversely distributed information, teams with pro-diversity views performed better than teams with pro-similarity views. In other words, merely espousing pro-diversity beliefs were insufficient to boost team performance in this study; cognitive diversity also needed to be present.

Papers that mathematically model groups have further supported the benefits of cognitive diversity in problem solving (Hong & Page, 2004) and forecasting tasks (Davis-Stober et al., 2015). For example, Davis-Stober and colleagues used a mathematical derivation of the mean squared error (MSE) to show that minimizing the covariance between forecasters' estimates reduces error in forecasts. Low covariance between forecasters' estimates is a form of cognitive diversity, as presumably when there is low covariance people would be using different cognitive strategies to produce estimates.

Although the methods and dependent variables in the mathematical modeling literature are very different from the approaches taken in most of the organizational literature, mathematical models are extremely informative as they support the importance of cognitive diversity for group tasks. In the present work, we employ both modeling and empirical work to explore social diversity and cognitive diversity's effects on performance.

Mechanisms. In the rare cases where social diversity improves group performance, how might those improvements arise? Several mechanisms have been suggested. The first, which has been called the "trait" method, suggests that people from different social categories have different knowledge, skills, and abilities that they bring to the group (McGrath et al., 1995; Sommers, 2006). Like pieces in a puzzle, their contributions can be combined to produce a superior outcome. Other researchers suggest that people from different social categories do not necessarily bring different skills or ideas, but their team members expect them to. These expectations and related team behaviors in turn elicit different contributions from team members, fulfilling the prophesy predicted by the "trait" method (Eagly & Wood, 1991; McGrath et al., 1995). Yet others suggest that diversity boosts performance via its effects on group processes like task conflict (Jehn et al., 1999; Pelled et al., 1999; Phillips et al., 2004), although they typically argue that surface demographic diversity is not helpful, only deeper, task-related diversity.

A fourth mechanism occurs via individual processing differences; people in diverse groups may individually reason differently than people in homogeneous groups. This idea is fairly well-supported. Racially diverse mock juries deliberate longer and recall facts more accurately than all-White mock juries, and this improvement is contributed not by Black members, as suggested by the "trait" theory, but by White members (Sommers, 2006). The author's interpretation was that when White jurors expected to deliberate with a diverse group, they processed information more systematically. The defendant in the mock trial was Black, however, so that study did not reveal whether racial diversity helps cognition for all tasks or only for those that are race-relevant. In subsequent work, Sommers, Warp & Mahoney (2008) found that people in racially diverse (vs. all-White) groups exhibited better memory and comprehension for race-relevant information, but not for race-irrelevant information.

Levine et al. (2014) had participants in a laboratory engage in a trading task. The market was either ethnically homogeneous or diverse. Although they worked individually, participants in homogeneous markets were less accurate in pricing than participants in diverse markets. The authors interpreted the better performance of participants in diverse markets as indicative of excess confidence and non-critical thinking present in people navigating homogeneous markets. Diversity may have made people more wary and cautious, leading to more careful thinking.

Also in support of the "individual" mechanism, Loyd, Wang, Phillips, and Lount (2013) found positive effects of political diversity on individual pre-meeting deliberation. When participants expected to solve a murder mystery in a politically diverse dyad as opposed to a homogeneous one, their pre-meeting written statements about the mystery exhibited more complexity and nuance in handling the facts. Relationship focus, or a need to maintain positive relationships, was established as a mediator. Individuals expecting to interact in diverse dyads exhibited better thinking because they had a lower relationship focus than people in homogeneous dyads. In other words, people in homogeneous dyads were more likely to believe that having a good relationship with their partner was more important than getting the facts right.

Although there is general supporting evidence for this mechanism, effects of social category diversity on reasoning have not been uniform. Sommers and colleagues (2008) also measured reasoning complexity found no reliable evidence that it was boosted by racial diversity. Other research has also failed to find a link between experimentally manipulated race diversity and individual integrative complexity (Antonio et al., 2004). They only found correlational results such that those who reported more inter-racial exposure exhibited more complexity in their thinking.

Implications for the present work. The work in this paper focuses both on “surface-level” diversity, specifically “social category diversity”, and “deep” diversity which corresponds to people’s estimates and biases on judgment tasks. The “surface” diversity must be linked to “deep” diversity, because when using crowds it is this deep diversity that helps boost accuracy.

Based on the emphasis on demographic diversity in previous work, most of the studies here measure demographic variables such as age, sex, and ethnicity, as well as less visible demographic variables such as educational attainment, political orientation, and religion. In line with previous theory, I expected each type of diversity to help crowd accuracy only to the extent to which that type of diversity was relevant to the judgment. For example, I thought political diversity should be particularly helpful on judgment tasks related to political questions.

As for the mechanism by which social diversity produces wiser crowds, only the “trait” mechanism described above is relevant. That is, if diversity helps, it will be because people bring different “puzzle pieces” to the table. Social diversity may lead to estimate diversity which can increase accuracy. Social diversity cannot improve performance via group processes because there are none when using crowds, and it cannot be via individual changes in cognition since participants have no expectation of interacting with group members, nor do they know anything about the people with whom they will be “combined” into a crowd.

It is also important to note that performance is subjectively assessed for the vast majority of the aforementioned research on diversity and performance. There are rarely gold-standard, “correct” solutions for the task at hand. Team performance in organizations is typically assessed according to the organization’s local methods of assessment, with good scores indicated by profitability or efficiency. In jury deliberations (e.g., Sommers, 2006), groups can correctly or incorrectly recall facts, and can correctly or incorrectly interpret the law (Ellsworth, 1989), but whether that recollection translates into proper judgment can be impossible to know. Some rare tasks are designed with correct answers (Phillips et al., 2004), but researchers have at times intentionally modified such tasks so that there is no correct answer (Loyd, Wang, Phillips, & Lount, 2013).

By contrast, the work in this dissertation is specifically interested in objective crowd performance, so questions must have a correct answer. Participants give numerical judgments in order to precisely state their views, and these are compared against a known correct value to determine performance. The implications of this point are further discussed in the introduction of Chapter III, but here I merely note this important distinction. Treating a problem as solvable versus subjective can greatly influence people’s judgments, especially when working in groups

(Stasser & Stewart, 1992), thus it is unknown how these diversity effects with “softer” performance measures shed light on diversity effects on concrete judgment tasks.

Three research questions

This dissertation addresses three main research questions related to the layperson’s practical use of aggregate judgments. First, a theoretical question is discussed: Under what conditions are socially diverse crowds more accurate than socially homogeneous crowds in numerical judgment tasks? A model described in Chapter II demonstrates that the answer depends on how much social factors bias judgment and also on where the truth lies relative to different groups’ judgments.

Second, how likely are those conditions to materialize in real life? The model presented in Chapter II sets the bar high – two conditions must both be true for social diversity to produce benefits above and beyond using homogeneous crowds. Seven studies in Chapter III examine people’s judgments on a wide variety of tasks. Results suggest that it is unlikely that commonly used social factors matter for crowd wisdom. Some homogeneous crowds are wiser than other homogeneous crowds, but diverse crowds are typically not wiser than the average homogeneous crowd.

Third, do laypeople in the United States have realistic expectations about the benefits of using socially diverse crowds? Two sets of studies explore this question in Chapter IV. The first studies examine whether people believe socially diverse crowds are wiser. They also gauge how much wiser people think those crowds are on a variety of judgment tasks. The second set of studies present behavioral implications of those beliefs. They test whether people prefer advice from diverse sources over homogeneous ones, and whether they are willing to pay more money for diverse crowd advice than for homogeneous crowd advice.

CHAPTER II

Conditions for Diversity Accuracy Gains

Modeling Social Diversity and Cognitive Diversity

I modeled the effect of crowd type, homogeneous or diverse, on accuracy. Various hypothetical "datasets" were generated following different parameters to test the effects of diversity on judgment accuracy. In each simulation, hypothetical estimates from different datasets were aggregated in a diverse or homogeneous manner, then compared against a hypothetical true value.

The model demonstrates three conditions that must be met in order for socially diverse aggregates to substantially outperform homogeneous aggregates. When all three conditions are met, aggregates of diverse estimates (from different groups) are more accurate than aggregates of homogeneous estimates (from the same group). First, the social characteristic of concern must systematically bias judgments. If, for example, men and women do not reliably make different types of judgments for the question at hand, then it would make no difference whether one employs aggregates of mixed gender or only one gender group. Second, the group difference must be large enough to have a meaningful impact on the bracketing rate and, by extension, accuracy. Third, the criterion – or correct answer – must lie between the average estimate of each group. In other words, the distribution of estimates from each social group must "bracket" the criterion (Larrick & Soll, 2006).

Method

I 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 group's guesses. As these parameters varied, the relative accuracy of diverse (vs. homogeneous) crowds varied. The manipulations, which I refer to as "effect size" and "bracketing", are described below.

Effect size. The model creates crowds by sampling from two normal distributions. Each distribution represents the population estimates from a homogeneous social group. For the sake of illustration, imagine one distribution represents estimates from Ohio State fans (OSU) and one distribution represents estimates from University of Michigan fans (UM). Theoretically they could be estimating anything, but for consistency, imagine 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 – for example, asking only UM fans for their estimates. Creating a diverse social crowd means drawing estimates from both distributions – asking 4 OSU fans and 4 UM fans for their estimates. I used aggregates of 8 based on evidence that wisdom of crowds effects are quite effective even for small crowds and accuracy benefits have diminishing returns as crowd size grows (Yaniv, 2004). Moreover, I wanted to simulate the size of a crowd a typical person might poll for advice – few people would have the resources or desire to poll over a handful of people when making a judgment.

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 = .1), the OSU and UM distributions will overlap a lot, almost to the point of being the same. Diverse crowds will on average be nearly indistinguishable from homogeneous crowds. If the effect of group membership on estimates is large (*r* = .8), the OSU and UM distributions will be highly distinct. Therefore, crowds from each homogeneous population will be distinct, and diverse crowds will on average yield an estimate that falls between the estimate of the two homogeneous groups.

Using R (R Core Team, 2016), four pairs of distributions were created to represent the following effect sizes: $r = .8$, $d = 2.6$, ($M_{\text{OSU}} = 58$ and $M_{\text{UM}} = 26$), $r = .6$, $d = 1.6$, ($M_{\text{OSU}} = 52$ and M_{UM}= 32), $r = .3$, $d = .6$, ($M_{\text{OSU}} = 46$ and $M_{\text{UM}} = 38$), $r = .1$, $d = .2$, ($M_{\text{OSU}} = 43$ and $M_{\text{UM}} = 41$). All standard deviations were 12.25, and the midpoint between both distributions' means is always 42[1]. Each distribution contained 10,000 "estimates" from that social group. Figure II.1 shows the distributions. To simulate crowds, I averaged 8 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.

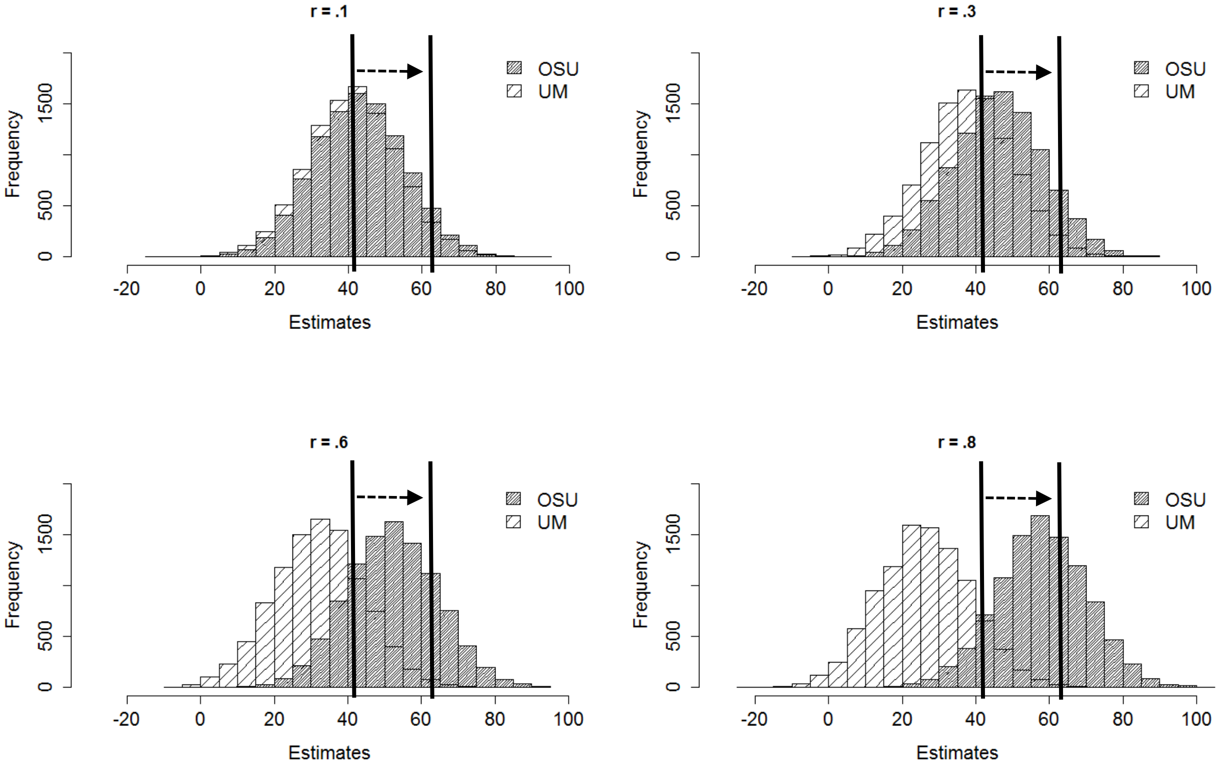
Bracketing. How well each crowd performs also depends on where the true value lies. Imagine UM's crowds predicted 38 points on average while OSU's crowds predicted 46 points. Diverse crowds on average predict the midpoint: 42. If OSU scores 7 points, the wisest crowd is a homogeneous one comprised of only UM fans. Diverse crowds have middling performance, and homogeneous OSU crowds would fare the worst. If OSU scores 42 points, the diverse crowd will be the most accurate, while UM crowds underestimate and OSU crowds overestimate the outcome. More generally, diverse crowds will be wiser than homogeneous crowds when the

[1] The numbers are arbitrary; other values can be used to create distributions at each effect size. I simply chose 42 because that is how many points OSU really scored in the 2015 match against Michigan, and 12.25 was roughly the standard deviation in a study we ran on football forecasts.

outcome lies near the midpoint between OSU's and UM's respective mean guesses. The farther away the truth is from the midpoint, the less of an advantage diversity offers.

In the model, accuracy is the absolute difference between each crowd's estimate and the true value (i.e., absolute error). Each crowd's performance is compared against varied hypothetical "true" values. In other words, crowd accuracy is re-computed at each hypothetical "true" value. Accuracy is first computed with the truth set to the midpoint, 42 (see bold black vertical line on the left of each plot, Figure II.1). It is then re-computed for each crowd at each effect size when the truth is set to 43, 44, 45, and so forth until 62, or 20 points away from the midpoint, towards M_{OSU} and ultimately passing it.

Figure II.1. Four Effect Sizes for Social Identity and Judgment



Note. This figure illustrates hypothetical distributions of two groups' estimates, Ohio State fans (OSU) and University of Michigan fans (UM). As labeled in the figure, each pair of distributions reflects a different effect size of group membership on judgment, from r = .1 to r = .8. Vertical black lines depict where the truth may lie relative to people's estimates in the simulations.

Results

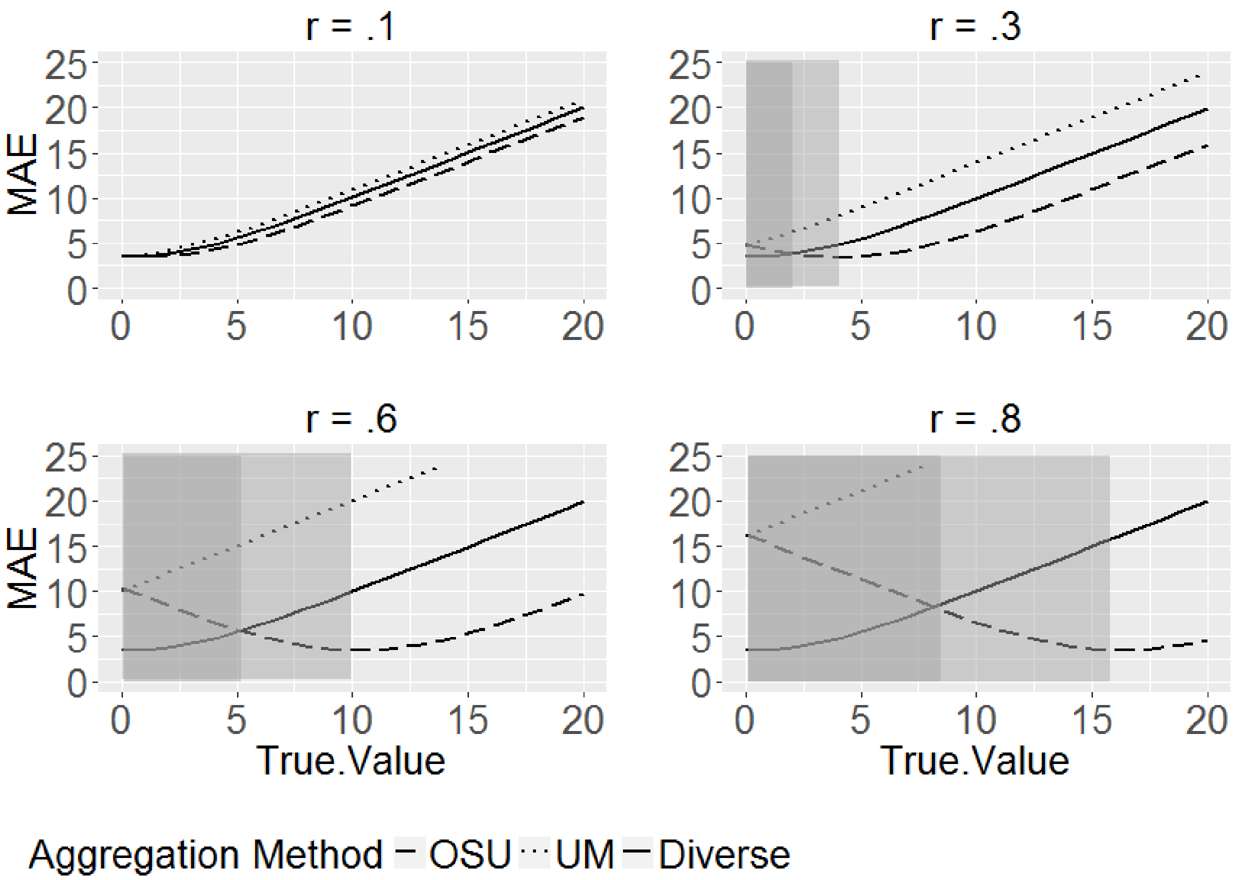
Figure II.2 models the accuracy of homogeneous and diverse crowds at different effect sizes for different degrees of bracketing. The y-axis denotes error; lower error indicates higher accuracy. The x-axis depicts variation in the true value, or bracketing. At the farthest point on the left (x = 0), the true value is set to lie at the midpoint between OSU and UM's guesses (refer to Figure II.1., left black line). It deviates 0 points from the midpoint. As x increases in 1-point increments, the true value shifts towards M_{OSU}, ultimately surpassing it (again see Figure II.1, the line trends right).

What happens to homogeneous crowd error depends on effect size and location of the truth. In Figure II.2, OSU crowd error decreases as the true value shifts towards M_{OSU}, but the error begins to increase when the true value shifts past and away from M_{OSU}. UM error continually increases as the true value shifts away from M_{UM}. OSU and UM crowd errors are similar at small effects (r = .1) and very different at large effects (r = .8).

Diverse crowds are sometimes more accurate than the homogeneous crowds. When social identity only weakly correlates with judgment (r = .1, Cohen's d = .2), diverse crowds of 8 are not notably more accurate than homogeneous crowds. Social identity must have a moderate to large effect on judgment for diverse crowds to outperform homogeneous crowds. When r = .3 (d = .6) and above, diverse crowds are the wisest crowds if the bracketing condition is met. (Note that the diversity gain is still very small at r = .3). But when the true value moves more than 50% closer to the mean of one group, the diversity advantage is lost, regardless of effect size. For example, at r = .6, M_{OSU} = 52 and M_{UM} = 32, and the midpoint, 42, is 10 points away from M_{OSU}.

Once the truth moves 5 points closer to M_{OSU} the diverse group ceases to be the most accurate, but it is still more accurate than the average homogeneous group. When the truth moves past M_{OSU} – that is, 10 points away from the midpoint in this case – the diverse crowd is as accurate as the average homogeneous crowd. It is less accurate than OSU and more accurate than UM. Shaded regions in the figure indicate diversity advantages over homogeneous crowds (see Figure II.2 caption).

Figure II.2. Accuracy as a Function of Crowd Type, Effect Size, and Bracketing



Note: On the x-axis, 0 indicates the truth is at the midpoint between the means of OSU and UM estimates (perfect “bracketing”). Increases on the axis indicate greater violation of the bracketing condition. Darker shaded regions indicate where diverse crowds outperform both homogeneous crowds. Lighter shaded regions indicate where diverse crowds outperform the average homogeneous crowd. Unshaded regions indicate where diverse crowds perform as well as the average homogeneous crowd.

In summary, the graphs all demonstrate that the accuracy of diverse groups decreases as the outcome value shifts away from the center of the two distributions. Small effects are especially susceptible to a loss of diversity benefits when bracketing is violated. Across all groups, when the bracketing condition is violated such that the criterion falls “outside” of the means of the two groups, the accuracy of diverse groups falls between the accuracy of the more accurate group and the less accurate group.

In Chapter III, a series of studies tests whether the diversity advantages observed in the model can be detected in real decision tasks. Socially diverse and homogeneous crowds are created from people’s estimates and their performance is compared. Consistent with the model, when conditions for a diversity advantage are not met, diverse crowds are merely as wise as homogeneous crowds.

CHAPTER III

Observed Diversity Accuracy Gains

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., Graham, Haidt, & Nosek, 2009) or questions for which the truth is subjective or cannot be known for certain (e.g., Kahan, Hoffman, & Braman, 2009). 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* = .21 (Richard, Bond, & Stokes-Zoota, 2003). 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 (.29) and are more empathetic (.37) than men (22). They are more likely to support the feminist movement (.39). But most gender effects were much smaller. For example, gender differences in social attribution were all smaller than .08. Female jurors are slightly harsher than male jurors in sexual assault cases (.16). Men are slightly more likely than women to dislike gays (.04 to .19). Boys are more competitive than girls by a small margin (.03), men have higher self-esteem than women (.06). Finally, race effects on judgment were also small: Whites report higher life satisfaction than African

Americans (.10). It is notable that of the few judgment questions studied, none were about matters of fact.

I am unaware of a body of research that systematically tests for social 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 constrained by reality and often diminish when there is accountability for one’s answers (Kunda, 1990; Lerner & Tetlock, 1999; Windschitl, Smith, Rose, & Krizan, 2010).

Studies III.1 – III.7

The present studies aim to address that gap in the literature by systematically looking for strong, significant correlations between social factors and people’s numerical judgments.

Method

For Studies 1-6, participants were recruited via Amazon’s Mechanical Turk (MTurk) platform. For Study 7, participants from diverse ethnic backgrounds were recruited via targeted e-mails from the university’s registrar’s office, followed by snowball sampling. Demographics from each study are summarized in Table III.1. Social factors were measured and correlated against numerical judgments in diverse domains. “Social diversity” referred to these demographic categories, as discussed in the Introduction: age, sex, ethnicity or cultural background, educational attainment, religion, and political orientation. I also examined fans from opposing teams in one study.

Table III.1. Demographic Information, Studies III.1-III.7

Study number	N	Mean age (*S.D.*)	% male	% White	Edu. attainment (*S.D.*)	Political orientation (*S.D.*)	Party identification			Religion	
							% Rep.	% Dem.	% Other	% Non	% Christian
1	51	38 (11.2)	51	86	3.53 (1.24)	3.59 (1.88)	-	-	-	33	41
2	198	33 (10.55)	61	72	3.71 (1.19)	3.44 (1.65)	25	60	15	43	45
3	564	37 (12.83)	38	74	3.67(1.14)	3.59 (1.59)	-	-	-	36	45
4	234	35.61 (11.91)	49	70	3.84(1.12)	3.47 (1.74)	22	52	27	38	47
5	205	35.51 (11.31)	36	77	3.79 (1.08)	-	-	-	-	-	-
6	226	39 (12.89)	38	82	3.64 (1.14)	3.62 (1.75)	-	-	-	40	42
7	202	27 (9.77)	32	50	4.21(1.35)	3.40 (1.65)	-	-	-	32	57

Note. Education was measured on a 6-point scale (1 = Some high school, 2 = Completed high school, 3 = Some college, 4 = Completed college, 5 = Some post-graduate education, 6 = Post-graduate degree). Political orientation was assessed on a 7-point scale (1 = Very Liberal, 7 = Very Conservative).

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 6 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 books. 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. Table III.2 summarizes the information for each study.

Table III.2. Study Summaries

Study	**Judgment task**	**Sample question**
1	Guess scores of 8 college football games in 2015	For the Ohio State Buckeyes vs the Michigan Wolverines game, what will be the final score? (Enter your guess for each team below by typing in numbers)
2	Forecast the percentage of votes that Republican and Democratic candidates will receive in Ohio and New Hampshire primaries, 2016 presidential election	What percentage of votes will Bernie Sanders receive in the New Hampshire primary election? (Answer must range between 0 and 100. Type in your number below, no % needed.)
3	Estimate the percentage of Americans that support a variety of polarizing political views	What percentage of Americans favor forming a federal database to track gun sales?
4	Predict what percentage of votes Hillary Clinton and Donald Trump will each receive in 10 states in 2016 United States presidential election.	What percentage of votes will Hillary Clinton receive in New York?

5	Guess the popularity rating of 24 books, previously tested among 50 MTurk users.	What was the average interest rating for this book? (*To Kill a Mockingbird*, by Harper Lee)
6	Estimate the likelihood of 40 diverse, domestic and international events occurring. Events derived from news stories	What is the likelihood of Myanmar's ruling government signing a peace agreement with rebel groups? What is the likelihood of California legalizing the recreational use of marijuana? What is the likelihood of Leonardo DiCaprio winning an Oscar?
7	Three judgment tasks: (a) predicting stock prices (b) predicting Olympic performance (c) predicting news events outcomes	(a) Click on where you predict this stock will be on June 19 2016. (Google price trend image) (b) In the 2016 Olympics, how many medals will Brazil win in Judo? (c) What is the likelihood that Chinese authorities will officially loosen new trade restrictions with North Korea?

Performance was lightly incentivized across most studies to promote recruitment and encourage attention. Top performers in the MTurk studies could earn between 2 and 10 dollars for their performance. For Study 7, 4 randomly chosen participants won $50 Amazon gift cards. The top-performing participant won a $100 Amazon gift card. Further details on the methods for each study are in Appendix A.

Results

Effect of social factor on judgment. To measure the effect of social identity on estimates, I correlated respondents' numerical judgments with social factors measured in each study. Where social factors were categorical, the categories were collapsed into two levels (e.g., religious = 1 and non-religious = 0, as opposed to 1 value per religion). In those dichotomous cases, point-biserial correlations were computed.

Hardly any strong relationships emerged when all judgments were correlated against all measured social variables. Out of 965 correlations, sixty percent were weaker than r ± .1 and 96% were weaker than r ± .2. Fewer than 1% of the correlations were .3 or stronger. These magnitudes suggest that for most of these questions it would be impossible for socially diverse crowds to outperform homogeneous crowds. All correlations significant at the *p* < .01 level are reported in Appendix B.

Homogeneous and diverse crowds. Despite the small observed association between social diversity and estimate diversity, could diverse crowds be wiser for the tasks that had the *largest* correlations between social identity and judgment? I compared the accuracy of socially homogeneous and diverse crowds from people's answers in these studies. To give diverse groups the best chance of outperforming homogeneous groups, I used tasks that showed biggest 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, their age, sex, political orientation, and whether they were religious. The observed *r* across questions for each task are shown in Table III.3.

From people's estimates, one thousand crowds were created by averaging 8 random people's estimates. Diverse crowds included 4 people from each social category (e.g., 4 liberals and 4 conservatives). Likewise, homogeneous crowds were created for each social category (e.g., 1,000 crowds including 8 randomly chosen liberals and 1,000 crowds including 8 randomly chosen conservatives). Individual and crowd accuracy were assessed by comparing mean absolute errors (MAE). People 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 – that is, the expected error of randomly choosing 1 homogeneous crowd.

It may be seen in Table III.3 that both homogeneous and diverse crowds typically performed better than the average individual across studies. But as Table III.3 also shows, crowd type didn't matter; diverse crowds usually 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: Ohio State did better than most people expected and Michigan did worse. 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 percentage point closer to the true outcome than the average homogeneous crowd comprised of only liberals or only conservatives. This amounted to a 10 percent decrease in error, from 8.91 to 7.99. No such advantage for diversity was observed in the other tasks.

It should be noted that participants in the sports study only made 2 target judgments – 1 per opposing team. The other teams were not Ohio or Michigan teams and so a bias was not present there. A better test of the bracketing condition for sports would involve multiple judgments as in Studies 2 through 7, because the failure to bracket in this single game may have simply been a fluke. In another study, I 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, with *r* not significantly differing from 0. Perhaps past games are not as compelling as an immediate upcoming game, or perhaps making multiple judgments reduces bias as people reason about the question repeatedly. Regardless of the reason, diverse crowds would not outperform homogeneous crowds in either sports study.

Table III.3. Accuracy of Individuals, Socially Homogeneous Crowds, and Diverse Crowds on Judgment Tasks

Task	Social identity	*r*	Average Individual Error (*S.D.*)		Average homog. crowd error	Average diverse crowd error
1	Team	0.36	14.49	(5.25)	13.6	13.52
2	Religiosity	0.13	14.46	(4.65)	11.85	11.79
3	Age	0.11	18.19	(6.05)	10.06	9.9
4	Political orientation	0.21	14.52	(6.70)	8.91	7.99
5	Age	0.1	2.24	(.57)	1.18	1.14
5	Sex	0.1	2.24	(.57)	1.07	1.06

Note. "Social identity" refers to the social factor that was most strongly associated with people's answers. The *r* is the average (absolute) correlation between that social category and answers. Average individual error is the average of absolute error across all questions in the task. The tasks were: (1) predict points in Ohio State vs Michigan game, (2) predict percentage of votes 8 presidential candidates would receive in 2 state primaries, (3) guess what percentage of Americans support each of 6 political statements, (4) predict what percentage of votes Clinton and Trump would each win in 10 states in 2016 presidential election (5) guess the popularity rating that 24 diverse books received in a previous study.

"Very homogeneous" and "very diverse" crowds. A stronger test of the diversity hypothesis would be to test the accuracy of "very homogeneous" groups against extremely 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. I examined such crowds' performance for the same tasks in Table III.3 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 2 or 3 social dimensions. Six homogeneous pools were created for each study based on factors that could reasonably be argued as relevant to the judgment (see Table III.4). 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. So 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, see note in Table III.4). 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*, I was able to create larger homogeneous pools, so the diverse pool included 100 participants.

Table III.4. Socially Homogeneous Groups

Task	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**
Predict percentage of votes 8 presidential candidates would receive in 2 state primaries	Random	White men, did not complete college	White women, completed college	Religious White Republicn.	Non-religious White Democts.	Liberal women under 40	Liberal non-Whites
Guess what percentage of Americans support each of 6 political statements	Random	White men, did not complete college	White women, completed college	Religious White Conservtv.	Non-religious White Liberals	Liberal women under 40	Liberal non-Whites
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 Republicn.	Non-religious White Democts.	Liberal women over 40	Liberal non-Whites
Guess the popularity rating that 24 diverse	Random	Men over 40	Men under 30	Women over 40	Women under 30	Ethnic minority women	White men

books received in a previous study

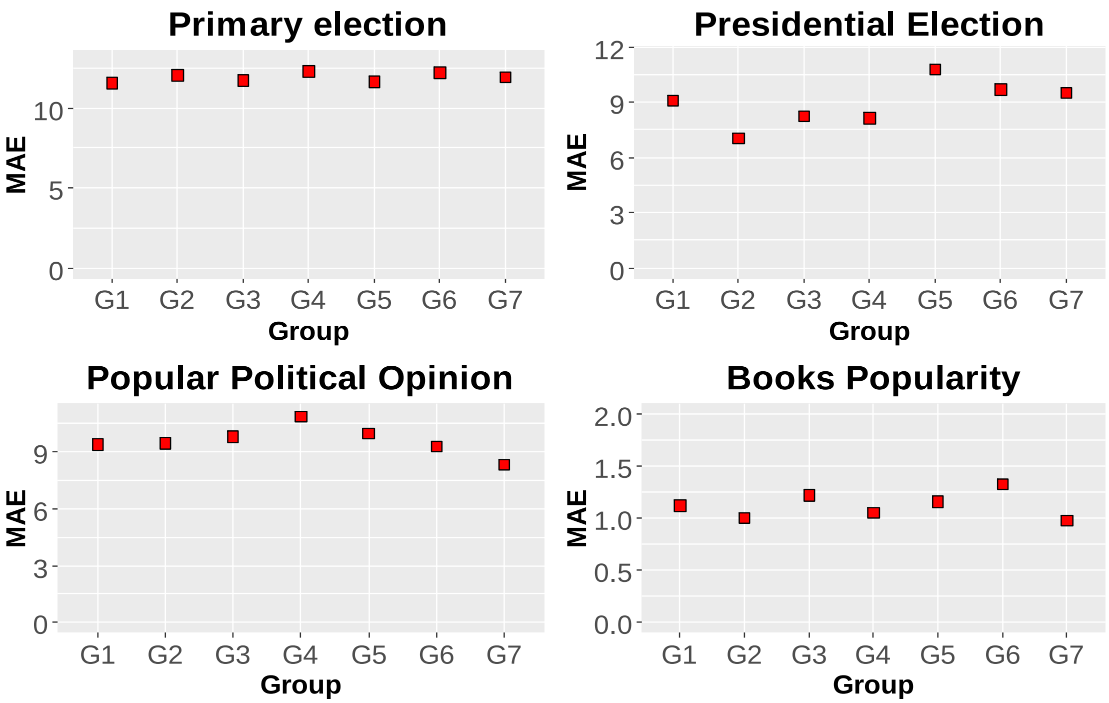
Note. All crowds except 2 were simulated from pools of at least 30 people. G2 for the book task (men over 40) was simulated from a pool of 22 men, and G6 (ethnic minority women) was sampled from a pool of 29 due to limited representation of those groups in the larger sample.

Homogeneous and diverse aggregates of 8 people were created by randomly sampling people 1,000 times. Accuracy was again assessed as MAE. Diverse MAE and homogeneous MAE were compared with the homogeneous groups averaged together.

Diverse crowds did not consistently perform much better than the average homogeneous crowd. For the primary election task, $MAE_{DIVERSE} = 11.56$ and $MAE_{HOMOG} = 11.96$. For the popular political opinion task, $MAE_{DIVERSE} = 9.40$ and $MAE_{HOMOG} = 9.61$. For the national presidential election task, $MAE_{DIVERSE} = 9.09$ and $MAE_{HOMOG} = 8.89$. These differences are minute – less than 1 point on a 101-point scale as participants estimated percentages. And for the book popularity task, both the diverse and homogeneous crowd MAE equaled 1.12.

Disaggregating the performance of homogeneous crowds reveals some fluctuation in crowd accuracy. Performance broken down by crowd type can be seen in Figure III.1. Given variation in homogeneous crowd performance, 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 doing relatively poorly. If one has no basis for choosing one homogeneous crowd over another, a diverse crowd is a “safe” choice. Nevertheless, homogeneous group performance does not vary substantially.

Figure III.1. Error of “Very Diverse” and “Very Homogeneous” Crowds



Note. G1 always represents the diverse crowds. G2 to G7 represent the homogeneous groups as described in Table III.4. 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 years old. In all graphs, the *y*-axis indicates error. Lower values mean higher accuracy on the task.

Discussion

As the model in Chapter II demonstrated, social diversity can theoretically make crowds wiser, but only to the extent to which it corresponds to diversity in estimates. Moreover, those diverse estimates must also bracket the true value in at least some realizations of the judgment outcome. The work in Chapter III suggests that small crowds can be wiser than individuals, but it often does not matter whether the crowd is homogeneous or diverse along demographic dimensions.

These results suggest several conclusions. First, for numerical judgments, socially homogeneous crowds may produce about as much estimate diversity as socially diverse crowds.

In the present research, 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. Our work supports scholarly and scientific exhortations to avoid stereotyping (Page, 2007). People might achieve this by realizing that, for numerical judgments about matters of fact, there is a lot of cognitive diversity within homogeneous social groups (Judd, Ryan, & Park, 1991). For example, people ought to consider that not all women think alike, not all liberals think alike, and so forth.

These results do not mean that *no* social factors correlate with any types of judgment. As I previously discussed, in many cases social identity can influence more subjective judgment. These results merely suggest that according to the operationalizations of social diversity in the present work, which are consistent with many popular conceptions of "diversity", there is only a weak correlation between demographic factors and numerical judgment. Any measure that effectively taps into people's signature cognitive biases would be a better proxy for cognitive diversity than the demographic variables that are all too commonly used.

Chapter IV addresses remaining questions about laypeople's expectations regarding social diversity and crowd wisdom. Do they expect diversity to be advantageous, and why? The studies also test to what extent people act on those expectations – are they more likely to choose diverse crowds when seeking advice on a judgment task, and are they willing to pay more for diverse crowds than homogeneous crowds?

CHAPTER IV

Lay Expectations of Diversity Gains

These studies explored whether and why people expect socially diverse groups to be more accurate than homogeneous groups in numerical judgment tasks. In the United States, social diversity is often promoted on practical grounds. Diverse groups are expected to outperform homogeneous groups on a variety of tasks (e.g., Galinsky et al., 2015). As discussed in Chapter I, in some contexts social diversity does make individual judgments more accurate (Levine et al., 2014), or it can lead to higher-quality deliberations that precede judgments (Loyd et al., 2013; Phillips et al., 2004; Sommers, 2006; Sommers et al., 2008). But studies conducted on organizational teams overall find no direct benefit of demographic, social category diversity on team performance (Webber & Donahue, 2001). Moreover, as Studies III.1-III.7 suggest, social diversity is unlikely to provide accuracy benefits for socially diverse crowds in numerical judgment tasks.

For the studies in Chapter IV, I hypothesized that people would generally think social diversity boosts aggregate accuracy (H1), partially because social diversity is increasingly viewed positively in American public discourse and there are several plausible arguments for why it may be beneficial for group performance, as reviewed in the Introduction. In addition to testing H1, I tested potential sources of inaccuracy. People might overestimate the effect size of social identity on judgment (H2a) or they might assume that estimates from different groups are likely to reliably bracket the true value (H2b). In other words, their lay theory about diversity benefits may unfold like this: “People of different social backgrounds have different ways of thinking (H2a) which will bias them towards different sides of the truth (H2b), so diverse groups are more accurate than homogeneous groups (H1) because their errors can cancel out.”

Behavioral outcomes of these beliefs were also tested. If people do expect socially diverse crowds to be wiser than homogeneous ones, are they more likely to choose diverse crowds when seeking advice on a judgment task? Does their choice depend not only on the diverse nature of the advice, but also on the cost of obtaining that advice? The third hypothesis states that if the cost of obtaining diverse and homogeneous advice is equal, people will prefer advice from diverse crowds when making numerical judgments (H3). The fourth hypothesis states that people will value the accuracy gains that they imagine for diverse advice, so they will be willing to pay more for diverse advice than for homogeneous advice (H4).

Studies IV.1-2 test people’s beliefs about the relative accuracy of diverse crowds (H1 – H2). In those studies, people imagine the average estimates that different social groups might give on a judgment task. They report how accurate homogeneous and diverse crowds might be for those particular tasks. Studies 1a, 1b, and 1c employ a correlational design while Study 2 experimentally manipulates bias and bracketing to test its effects on people’s beliefs about diverse crowds. Studies IV.3-IV.5 examine advice choice and willingness to pay for advice to examine the behavioral implications of diversity beliefs (H3 and H4).

Study IV.1

In Studies 1a-1c, people guessed the numerical judgments of people from particular social groups to test H2a and H2b. The questions measured people’s *imagined* effect of social identity on judgment. They were also asked what types of crowds they thought would be most accurate for the judgment at hand, to test H1. If they thought diverse crowds would be more accurate than homogeneous ones, they also indicated by how much.

For Studies 1a and 1b, I simulated homogeneous and diverse crowds created from people's imagined estimates. For example, I took what they *thought* liberals and conservatives would predict for an election and treated them as real estimates. I then examined how accurate these imagined homogeneous and diverse crowds were in order to see if people's inferences from their assumptions were justified. In other words, if the bias that people imagined were real, then would diverse crowds be wiser than homogeneous crowds? The questions across these studies are summarized in Table IV.1.

Table IV.1. Summary of Studies IV.1a – IV.1c

Study	Domain	Groups	Target judgment	Sample Questions
1a	Politics	Conservatives vs. liberals	Percentage of votes that Hillary Clinton and Ted Cruz got in their respective primaries in Ohio	What outcome do you think **conservatives (liberals)** predicted, on average, for **Hillary**?
1a	Sports	Ohio State University football fans vs. University of Michigan football fans	Number of points Ohio State and Michigan each scored in their 2015 game	How many points do you think **Ohio State (Michigan) fans** predicted, on average, for **Michigan**?
1b	Feminine music show	Men vs. women	Attendance at a local music show featuring a young woman playing acoustic folk music	Out of all the festival attendees we poll today, what percentage will plan to attend this event? What prediction do you think **women** (**men**), on average, would make for question 3?
1c	Art fair	Men, women, out of town visitors, participating artists	How many artists were participating in the 2016 State St. art fair in Ann Arbor, MI	How many artists do you think are participating in the State Street 2016 art fair? How would **men** (**women/ out of town visitors/ participating artists**), on average, answer question 1?

Note. The "groups" column describes groups for which participants guessed their average answers.

Method, 1a

In Study 1a, (N = 201, M_{age} = 36.41, *S.D.* = 12.97, 40% male), MTurk participants answered questions related to politics and sports. Thirty two percent completed some college, 34% completed college, 21% completed some or all of a graduate degree. Fifty one percent identified as liberal to some degree, 24% were moderate, and 25% were conservative to some degree. The sample was 72% White, 12% Black, 6% East Asian, 5% Latino(a), and 6% mixed or "Other".

Imagined group estimates. Respondents' first task was to guess people's predictions from previous studies. For each question, participants were given the actual outcome and then guessed the mean predictions of different groups. For the politically themed questions, respondents guessed what percentage of votes conservatives and liberals predicted for Hillary Clinton in Ohio's primary in 2016. They also guessed what percentage of votes each group predicted for Ted Cruz. (These candidates were chosen because at the time of the study, both were faring relatively well in the presidential race and Cruz seemed to be a slightly less controversial candidate than Donald Trump.)

For the sports themed questions, respondents guessed Ohio State and Michigan fans' predictions of Ohio State's score in the 2015 Ohio State/Michigan football game. They also guessed each group's prediction for Michigan's score. All participants completed political and sports questions, with the order of the themes counterbalanced and with randomized questions within each condition. Responses were typed into an open-ended text box.

Choice of optimal crowd. After guessing, participants were told that averaging estimates in different ways can often yield relatively accurate predictions. I asked them what kind of strategy might be more effective in each context. For example, in the politically themed portion, participants could choose one of the following: (a) “Overall, averaging guesses from similar people (only one political party) will be most accurate,” (b) “Overall, averaging guesses from diverse people (different political parties) will be most accurate,” or (c) “Overall, the performance of similar vs diverse groups will be about the same.” Order was randomized across options.

Estimated accuracy gains. If participants chose the homogeneous or diverse option, they were then asked by how much that option would reduce error as compared to the other option. Responses were given on a scale with 20% increments: “It’s 90% closer to the truth,” “It’s 70% closer to the truth,” and so on through 10%.

I quizzed participants on the instructions after the second portion to check how much they understood them. Eighty four percent of respondents correctly indicated that, among other things, they were asked to “guess what percentage of votes liberals predicted for Hillary.” Incorrect answers (“guess the number of votes Hillary got” and “guess what percentage of liberals voted for Hillary”) were chosen by 5% and 11% of participants, respectively. Filtering out participants who did not correctly answer this question did not alter estimates in any substantial way, so the analyses reported here do not use this filter.

Finally, participants completed a brief demographic questionnaire indicating their age, sex, ethnicity, educational attainment, and political ideology.

Method, 1b

In Study 1b, participants encountered a less adversarial domain – namely, they estimated the predictions of men and women for attendance at a young woman’s folk performance. Experimenters approached people attending the festival and asked them to take an event planning survey. Sixty four people (49% male, M_{age} = 33.16, *S.D.* = 15.20) completed the survey in exchange for a $1 coupon for a local coffee shop. They first read a brief event description taken from the official event website. The event was a free music performance scheduled for that Friday. I chose it specifically because it appeared feminine and appealing to women as it featured a young, female artist. On the survey, a flower embellished the event description to reinforce the feminine aspect of the event. Appealing to one demographic makes it quite reasonable to believe that men and women would give different estimates.

After reading the description, participants indicated their own interest in the event (1 = Not at all interested, 4 = Extremely interested) and whether they planned on going (1 = Definitely won't, 4 = Definitely will). Then, they guessed what percentage of our respondents would plan to attend the event.

Imagined group estimates. In counterbalanced order, participants then guessed what the average attendance estimate would be from our female respondents and from our male respondents.

Choice of optimal crowd. They then guessed what type of group would be most accurate. They read: "We want to get an accurate prediction of turnout for the event. We'll ask people to guess what the turnout will be, and average their answers. To be most accurate, should we ask: (a) Just women for their estimates, (b) Just men for their estimates, (c) Both women and men for their estimates, (d) All of the above will be equally accurate."

Estimated accuracy gains. They were then asked to indicate how much more accurate their selection would be: "If you chose a, b, or c (question 6), how much more accurate do you think that group will be compared to the other choices? (a) 10% closer to the true number, (b)

30% closer, (c) 50% closer, (d) 70% closer, (e) 90% closer to the true number." After giving their answer, participants indicated their age and sex.

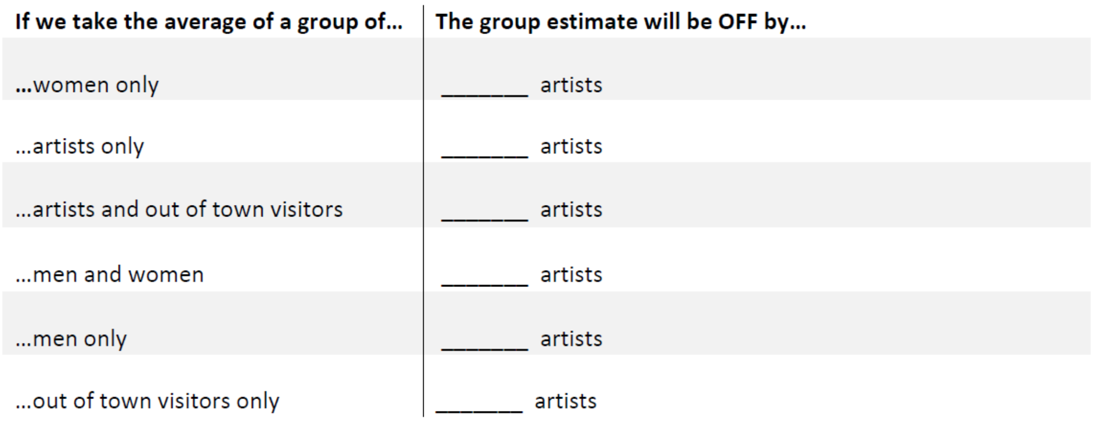
Method, 1c

Sixty six art fair attendees participated in exchange for a $1 coupon to a local coffee shop. One person gave mathematically impossible error estimates (i.e., error estimates were larger than target estimates), so responses from 65 participants were included in analyses (M_{age} = 48.03, *S.D.* = 17.07, 40% were male, 22% were local residents). Everyone first answered the question, "How many artists do you think are participating in the State Street 2016 art fair?"

Imagined group estimates. They then answered four questions about how four different groups would have answered question one: "How would (**participating artists/ out of town visitors/ men/ women**), on average, answer question 1?" They answered by filling in a blank: "_______ artists."

Choice of optimal crowd. In this study, rather than choosing the wisest crowd, participants indicated by how much each crowd type would err. The instructions stated: "Below, take a guess at how accurate we'd be if we averaged guesses from each of the following groups." They then indicated their error estimates in table form as shown in Figure IV.1. The researcher answered any questions that participants had while they completed the surveys.

Figure IV.1. Answer Format, Study 1c



If we take the average of a group of…	The group estimate will be OFF by…
…women only	_______ artists
…artists only	_______ artists
…artists and out of town visitors	_______ artists
…men and women	_______ artists
…men only	_______ artists
…out of town visitors only	_______ artists

Note. To make the task as clear as possible, we asked participants to answer the question in table format as depicted here.

Finally, participants indicated their age, sex, and whether they lived in the town that was hosting the art fair.

Results

Diverse groups are best (H1). H1 was supported for the domains in which people had lay theories about bias – specifically, politics, sports, and women's folk shows. Most people thought diverse groups would be the optimal crowd. In Study 1a, 58% of people thought diverse aggregates would be the most accurate in political domain and 56% of people thought diverse aggregates would be the most accurate in the sports domain. In Study 1b, 84% of people thought diverse aggregates would be more accurate than homogeneous aggregates in predicting concert attendance.

Estimated accuracy gains were high. Participants who thought that diverse crowds were optimal said such crowds would decrease error by 57% (political) and 55% (sports) as compared to polling a homogeneous group. In Study 1b, people who thought that diverse crowds were optimal said such crowds would decrease error by 42.92% on average (*S.D.* = 23.15).

In Study 1c, I did not ask participants to choose which type of group would be most accurate in estimating number of artists. Instead, I asked them to estimate the error of diverse and homogeneous groups. People did *not* expect diverse groups to have lower error than homogeneous groups. As will be seen below, people did not expect different social groups to be biased, unlike in Studies 1a and 1b. Thus, they perceived no advantage of diverse crowds over homogeneous crowds.

Imagined group estimates (H2). People's guesses about different groups' predictions are presented in Table IV.2. H2a was supported in that people typically overestimated the effect of social identity on judgment. In political judgments, for Ted Cruz, people thought conservatives would make higher performance predictions than liberals. For Hillary, people thought conservatives would make lower performance predictions than liberals. As for football score estimates, people thought Ohio State (OSU) fans would give higher estimates than Michigan (UM) fans when predicting OSU's score. The reverse was true for predicting UM's score. When guessing the mean estimates of men and women for attendance to the folk music performance, people thought women would give higher estimates of attendance than men. However, for the number of artists at the art fair, people did not think that estimates from different groups would differ, $F(3, 54) = 1.43$, $p = .244$. For all other studies, the r value was computed by treating people's imagined mean estimates as if they were real, and thus r represents the point-biserial correlation between group membership and judgment.

Table IV.2. People's Estimates of Other People's Predictions

Estimate	**True Value**		**Mean**	***S.D.***	***t***	***df***	***p***	***d***	r_i	r_t
Cruz	13.1	Conservative	38.94	21.47	10.26	196	<.001	.99	.45***	.05
		Liberal	20.88	14.15						
Hillary	56.5	Conservative	38.51	19.87	8.90	195	<.001	.94	.43***	.02
		Liberal	56.66	18.67						

Ohio State	42	OSU fans	39.21	13.43	18.65	198	<.001	1.63	.63***	.38**
		UM fans	19.38	10.70						
Michigan	13	OSU fans	15.95	11.76	14.19	199	<.001	1.23	.52***	.34*
		UM fans	32.09	14.42						
Music show	42	Women	46.7	23.10	5.59	61	<.001	.60	.29**	.13
		Men	33.32	21.10						
Art fair	351	Men	311.89	332.77	-	-	*n.s.*	-	-	-
		Women	288.05	395.45						
		Artists	285.14	314.47						
		Visitors	370.11	610.16						

Note: Means reflect what people *thought*, on average, other people estimated in each of the judgment tasks. For example, participants thought that conservatives predicted Cruz's performance to be 38.95% of the vote. They thought that liberals predicted Cruz's performance to be 14.15% of the vote. The r_i column indicates the imaginary effect of social identity on judgment, while r_t indicates the observed effect from previous studies. *, **, and *** indicate that r was significantly different from 0 at $p < .05$, $p < .01$, and $p < .001$ respectively.

Hypothesis 2b was not consistently supported; sometimes mean estimates bracketed the true value and sometimes they did not. In Study 1a, for estimates of Cruz' performance, the means (and medians) of imagined conservative and liberal predictions do not bracket the true value of 13.1. For estimates of Hillary's performance, the means and medians barely bracket the true outcome of 56.5. For football scores, in neither case do the estimates of OSU fans and UM fans bracket the true outcome value. For Study 1b, however, people did imagine men and women's estimates to bracket the criterion. In Study 1c only visitors overestimate the truth, all others underestimate. (Note that in Studies 1b and 1c, people were not given the true value.)

Estimated accuracy gains. In Studies 1a and 1b, people reported by what percent diverse crowds would decrease error over homogeneous crowds (if they had indeed chosen diverse crowds as optimal). Those expectations are reported here. Additionally, homogeneous and diverse crowds of 8 were simulated from people's *imagined* mean estimates of how each group would respond. One thousand crowds of each type were created. The accuracy of these simulated diverse (vs. homogeneous) crowds will reveal whether people's estimated accuracy gains are reasonably derived from their imaginary estimates, or whether they are overblown. To summarize, I treated their imagined estimates as if they were real and addressed the question, "If people's assumptions were true, then would their inferences about diverse crowd performance be accurate?"

For Study 1a's sports questions, people who chose diverse crowds thought that they would decrease error over homogeneous ones by 55% on average (*S.D.* = 21.05). The simulated crowd accuracy based on their imagined mean estimates was as follows: MAE_{HOMOG} = 13.57[2] and MAE_{DIV} = 12.56 for OSU's points, and MAE_{HOMOG} = 11.39 and MAE_{DIV} = 11.12 for UM's points. Therefore, if people's assumptions were true, diverse crowds *would* have lower error than homogeneous crowds. But, the amount of error reduction would be negligible because people did not imagine much bracketing. In short, an inference problem is also present: people overestimated how much accuracy diversity would buy them even if their beliefs about bias had been correct.

For Study 1a's political questions, people who chose diverse crowds thought that they would decrease error over homogeneous ones by 57% on average (*S.D.* = 19.22). The crowd accuracy based on their imagined mean estimates was as follows: MAE_{HOMOG} = 16.78 and MAE_{DIV} = 16.64 for Cruz's performance, and MAE_{HOMOG} = 11.71 and MAE_{DIV} = 9.82 for Hillary's performance. People again overestimated by what margin diverse groups would decrease error – for Hillary's performance, diverse crowds would have offered a 16% decrease in error, not 57%.

For the concert attendance task in Study 1b, people who chose diverse crowds thought that they would decrease error over homogeneous ones by 43% on average (*S.D.* = 23.15). Simulations revealed that MAE_{HOMOG} = 8.32, and MAE_{DIV} = 6.17, a decrease of about 25%. Here,

[2] For all of these results, error is collapsed across the homogeneous crowds.

people overestimated the degree of diversity benefits, but to a lesser extent because their imagined estimates bracketed the truth more evenly than imagined estimates in the other tasks.

Discussion

For three out of four judgment tasks, people expected diverse crowds to be more accurate than homogeneous crowds (H1). For those same studies people overestimated how much group membership predicts judgment (H2a). They also expected bracketing in some, but not all, cases (H2b). But when people did not expect group membership to influence judgment (Study 1c), they did not expect diverse crowds to be wiser than homogeneous ones.

Thus, people only seem to expect social diversity benefits when they think the conditions for those benefits have been met to some degree. When they think people are biased, they expect diversity to trump homogeneity. When they think that people are not biased in a particular domain, they do not expect diverse crowds to be wiser than homogeneous ones.

It is notable that people largely overestimated the differences between social groups' estimates. Some simulations tested for diversity advantages assuming that people's imagined beliefs were in fact true. Diverse crowds did consistently outperform homogeneous ones, but not nearly to the degree that people expected them to. Thus, although false beliefs about diversity seem to stem from wrong premises about bias, there is an inference error of degree as well.

There are some limitations to this series of studies. First, there was some inconsistency in the bracketing results. Overall, people's estimates for two social groups did not always bracket the true answer, even when they were given the true answer before providing their estimates. Another weakness is that I did not measure the dispersion that people may have imagined for each group's estimates. For example, I did not ask people to guess the standard deviation for conservative and liberal guesses. However, in another study not reported here I collected measures of dispersion and found that people both overestimate the difference between groups' estimates and underestimate the variance within groups' estimates. Finally, people in Study 1c gave error estimates for each crowd instead of choosing the wisest crowd like in Studies 1a and 1b. This different response modality could have influenced the results. Study IV.2 more rigorously tested whether people expect diversity benefits only when the bias and bracketing conditions are met.

Study IV.2

This study experimentally manipulated effect size and bracketing to test the effects of these conditions on people's judgments about diverse crowd accuracy. In the study, people read hypothetical estimates from two different groups for a variety of judgment tasks. They also saw the true outcome, and then indicated whether sampling from only one of those groups (homogeneous) or both of those groups (diverse) was most likely to yield the most accurate prediction. It was hypothesized that when (1) groups were biased, with a visible effect of demographic membership on judgment, and (2) when estimate distributions bracketed the truth, people would be more likely to think diverse groups outperform homogeneous groups than when those conditions were not met.

Method

One hundred and fifteen participants from the introductory psychology subject pool participated for course credit. The mean age was 18.53, *S.D.* = 1. Forty three percent were men, and the ethnic composition was: 7% Black, 14% Asian, 5% Latino(a), 5% Mixed or Other, 9% Other Asian or Pacific Islander, 60% White, non-Hispanic). Forty nine percent of participants had previously taken a statistics course.

Students participated in study sessions in groups of up to 5 people. At the start of each session, the experimenter gave a brief lesson on how to interpret histograms. The lesson was provided to make sure that everyone would be able to understand the questions in the study, but it provided no information on wisdom of crowds or judgment aggregation. The lesson PowerPoint slides included example histograms and the experimenter explained what the axes represented in each example. Participants identified the most frequent values across some example histograms, and they also interpreted an example with two distributions on the same plot. Finally, they were given a preview of the types of questions they would answer in the survey. Participants then proceeded to answer the survey questions on paper.

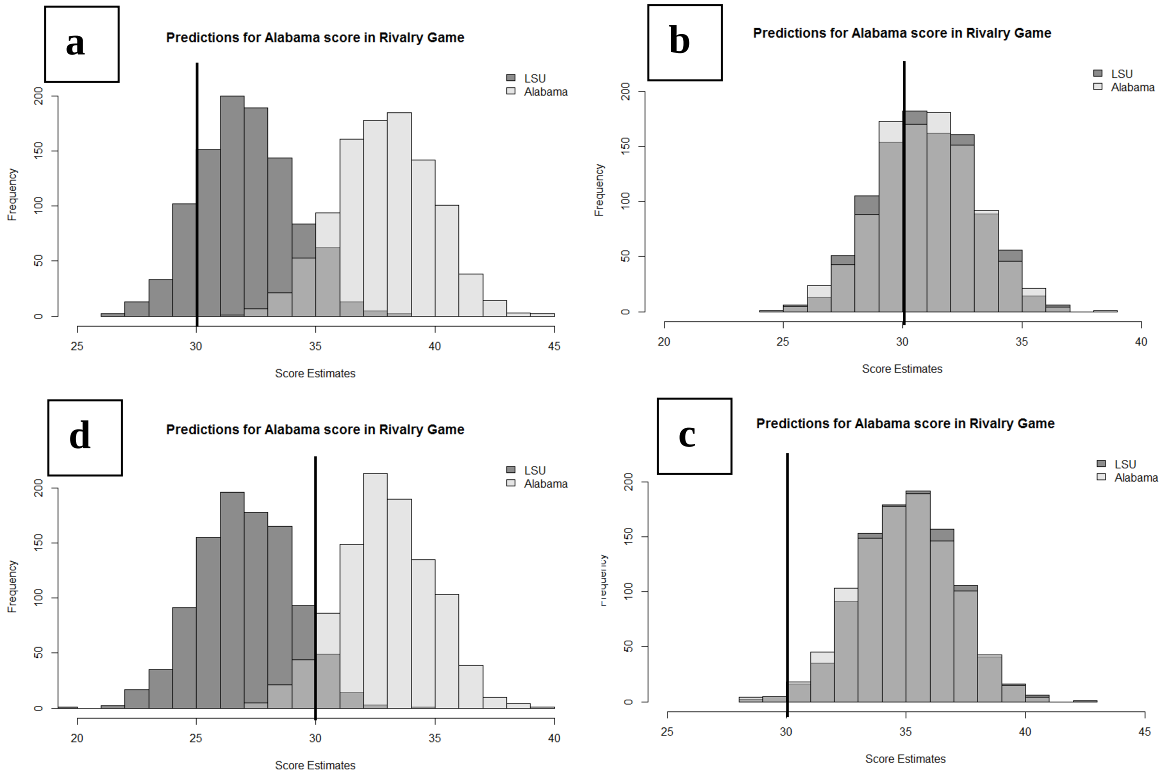
The task contained 12 scenarios and questions like the one below. In the example, the team people cheer for clearly biases their estimates, and their estimate averages bracket the truth:

> Last year, Alabama played LSU. Alabama won with a total of 30 points. A random sample of students from Alabama and LSU were polled before the game. They guessed how many points Alabama would score.
>
> On average, Alabama students predicted that Alabama would get 33 points while LSU predicted 27 points. The graph shows the distributions of each group's guesses.

Participants then saw a corresponding image that illustrated the distributions of each group's guesses (see 4 examples in Figure IV.2). They then were told: "Imagine you will poll 10 people and average their guesses together. If you want to be as accurate as possible, which approach would you take below? (clearly circle one)" and were given four options corresponding to the question. For the example question above, they would see these options:

> (a) Sample of only Alabama students
> (b) Sample of only LSU students
> (c) Mixed sample of Alabama and LSU students
> (d) All of the approaches will be equally accurate.

Figure IV.2. Bias by Bracketing (2 X 2) Design



Note. In clock-wise order, they represent (a) bias, no bracketing, (b) no bias, bracketing, (c) no bias, no bracketing, and (d) bias and bracketing. Diverse crowds are optimal for (d) only.

Across 12 questions in a within-subjects 2 X 2 design, the presence (absence) of judgment bias was manipulated, as was the presence (absence) of bracketing. There were three questions of each type. To manipulate bias, participants scenarios in which estimates from the two groups were *different*, on average (bias condition) or the *same*, on average (no bias condition). To manipulate bracketing, the true value lay approximately at the midpoint of all estimates (bracketing condition) or it lay offset from the midpoint (no bracketing condition). Figure IV.2 shows the visual representation of each cell in the design.

Three domains were used in this study: sports, politics, and entertainment. Four questions were created for each domain. Within a domain, there was one question for each cell of the 2 X 2 manipulation; that is, 1 question with bias and bracketing, 1 question without bias and with bracketing, 1 question with bias and without bracketing, and 1 question without either bias or bracketing. Diverse crowds should only be chosen as the most accurate for the first type of question (with bias and bracketing).

Four versions of the survey were created using a Latin square approach to ensure that across all versions, each specific question was evenly paired with each variation of the manipulation. In each version, sports questions were presented first, followed by political questions and then by entertainment questions. Four more versions were then created to present the domains in an alternate order: politics, sports, and entertainment.

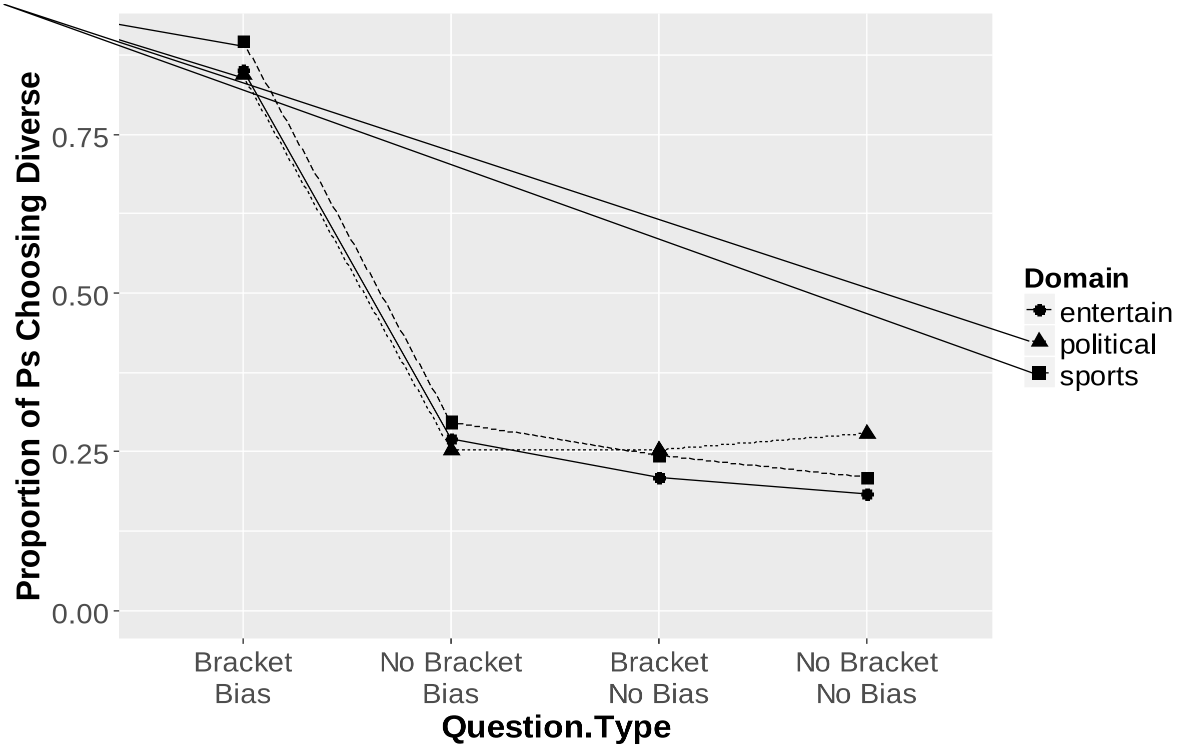
Surveys were randomly distributed to participants in each session. After answering the questions, participants indicated their age, sex, and ethnicity. They also indicated whether they had previously taken a statistics course (yes or no).

Results

The main question was whether participants were more likely to choose the diverse crowd as optimal when both necessary conditions were met as opposed to when they were not. Thus for each question, participants' responses were coded as choosing the diverse option (1) or choosing another option (0). A logistic regression examined the effects of bracketing (present vs. absent), bias (present vs. absent), and domain on participants' choices. Importantly, the interaction of effect size and bias was tested. The model allowed random intercepts for participants due to repeated measurements. The model failed to converge using all of these parameters, so a separate regression was conducted for each domain.

Across domains, most participants correctly chose the diverse crowd as optimal when both bias and bracketing conditions were met. When one or neither condition was met, the number of participants choosing diverse crowds was reduced to about chance level (25%). Figure IV.3 shows the proportion of participants choosing diverse crowds for each question type across the 3 domains. In the 3 regressions, there was never a main effect of bracketing. Only for entertainment was there a main effect of bias (*OR* = 2.35, *Z* = 2, *p* = .045) such that people were about twice as like to choose diverse groups (vs. non diverse groups) when there was bias than when there was not. More importantly, the interaction was significant for all three domains, indicating that choosing a diverse crowd depended on both bracketing and bias (Political: *OR* = 41.55, estimate = 3.73, *Z* = 6.63, *p* < .001; Sports: *OR* = 67.74, estimate = 4.22, *Z* = 6.31, *p* < .001; Entertainment: *OR* = 90.34, estimate = 4.51, *Z* = 5.91, *p* < .001).

Figure IV.3. Effect of Bias and Bracketing on Diverse Choice



Note. For the 4 conditions in this study (x-axis), participants were most likely to choose the diverse crowd as optimal in the "Bracket Bias" condition as opposed to the other 3. This reflects normative thinking, as diverse crowds are only beneficial for cases where the social groups are differentially biased in their thinking and the truth lies near the center of people's estimate distributions.

To break down the interaction, variables were coded in the following manner: questions for which both conditions were met were coded as 1, and each of the 3 remaining questions were coded as -1/3. This allowed me to contrast the question for which diverse crowds were optimal against all other questions for which diverse crowds were not optimal.

In the political logistic regression, meeting both conditions meant the odds of choosing (vs. not choosing) diverse crowds was about 14 times greater than for questions in which both conditions were not met (*OR* = 13.77, estimate = 2.62, *Z* = 8.87, *p* < .001). For the sports logistic regression, meeting both conditions meant the odds of choosing (vs. not choosing) diverse crowds was about 37 times greater than when both condition were not met (*OR* = 37.21, estimate = 3.62, *Z* = 9.05, *p* < .001). When both conditions were met in the entertainment domain, the odds of choosing (vs. not choosing) diverse crowds was 47 times greater than when both conditions were not met (*OR* = 47.34, estimate = 3.86, *Z* = 8.34, *p* < .001).

Several measures of accuracy were also analyzed, including proportion correct, sensitivity, specificity, and ΔP (all measures described in Yates, 1990). Proportion correct indicates how often people were correct in identifying cases where diverse crowds were (or were not) optimal crowds [3]. The mean proportion correct was .78 (*S.D.* = .23), suggesting that participants on average were correct in their judgments almost 80% of the time.

The sensitivity and specificity of people's judgments were also examined. Sensitivity indexes to what extent a target is indeed present when people say that it is. High values indicate

[3] People's true positives and true negatives were summed and divided by 12, the total number of judgments. "True positives" refer to instances where participants correctly said diverse crowds were best. "True negatives" refer to instances in which participants correctly did not say that diverse crowds were best.

the person’s answers largely included true positives, while low values indicate that a person’s answers largely include false positives. Specificity indexes to what extent people are correct when they say a target is absent. High values indicate the person’s answers largely included true negatives, while low values indicate many false negatives (that is, people *missed* the presence of the target).

Sensitivity is computed as a participant’s true positives divided by the total number of positives that s/he indicated. Mean sensitivity was .65 (*S.D.* = .19), so for every three times people said diversity was optimal, it was truly optimal for 2. Specificity is computed as a participant’s true negatives divided by the total number of negatives that s/he indicated. Mean specificity was .76 (*S.D.* = .31), so for every four times people said diversity was not optimal, it was truly not optimal for about 3.

Finally, ΔP was computed for each participant. This is a contingency statistic reflecting the association between people’s judgments and the truth. It is computed by subtracting the false positive rate from the true positive rate (Yates, 1990). Scores of 1 and -1 indicate a perfect relationship between judgments and the truth, and 0 indicates no relationship between judgments and the truth. The average ΔP value was .40 (*S.D.* = .35), and it was significantly different from 0, $t(114) = 12.38$, $p < .001$. The positive sign reflects that people’s judgments positively covaried with the truth; people were relatively likely to say diverse crowds were optimal when they were truly optimal, and relatively unlikely to say diverse crowds were optimal when they were not truly optimal. Whether a participant had taken a statistics course before or not had no association with ΔP scores, $t(114) < 1$.

Discussion

When bias and bracketing are experimentally manipulated, people's judgments about diversity advantages are sensitive to both. They were far more likely to say that diverse crowds were optimal when both bracketing and bias were present than when one or both were missing. The pattern held across three domains.

These results suggest that when people expect diversity advantages, it is because they are assuming some degree of bias and some degree of bracketing. When those factors are manipulated, people are reasonably accurate in determining when to expect diverse crowds to be optimal. They are not perfect, however; they show some degree of Type I and Type II error. As Figure IV.3 shows, when the necessary conditions are not met about 25% of people still choose diverse crowds as optimal, and when both conditions are met, fewer than 100% of people choose diverse crowds.

Study IV.3

Study IV.3 tests whether people prefer advice from a diverse crowd over advice from a homogeneous crowd. Although people in Study 1 reported thinking that diverse crowds are wiser than homogeneous ones, this does not necessarily translate into behavior. It is possible that people do not place much value on the perceived accuracy gains that diverse crowds offer, or it is possible that people prefer advice from homogeneous groups. In information seeking tasks, people often prefer information that is consistent with their own views and biases (Hart et al., 2009), so this may lead people to prefer information from homogeneous crowds comprised of similar others. Conversely, people may think that homogeneous crowds comprised of dis-similar others are more helpful in that they can serve as a counterbalance against one's own bias. Thus a behavioral measure, choice, was used to test H3: People will prefer to see advice from diverse crowds when making estimates in judgment tasks.

The study was structured to measure the preference for diversity conservatively. Specifically, people had the option to choose to see answers from three kinds of crowds: a diverse crowd of two people (1 liberal and 1 conservative) or a homogeneous crowd of three people (3 liberals or 3 conservatives). In other work, I have demonstrated that homogeneous crowds of 3 are often just as wise or *wiser* than diverse crowds of 2 (de Oliveira & Larrick, in preparation). Therefore, normatively speaking people should not show a preference for diverse crowds.

Method

People were approached in public spaces near a large university and invited to participate in a short advice preference study. One hundred and forty-six people participated (M_{age} = 20.72, *S.D.* = 4.83; 79 men, 64 women, 1 non-binary and 3 did not report sex). They were told they would answer some trivia questions on paper and could earn a $1 off coupon for a local coffee shop if their answers were accurate enough. They were also told that to help with the task, they should choose what type of advice would be useful to them. The advice consisted of answers from people who previously took the task.

Participants chose between seeing answers from 1 conservative and 1 liberal, 3 conservatives, or 3 liberals. Four versions of the survey rotated the order of diverse (vs. homogeneous) advice first, as well as whether conservatives or liberals were first. Participants received, in fact, about the same advice regardless of their choice options, only there were 2 or 3 columns depending on the size of the group they chose.

After making their selection and receiving a sheet of paper with the previous answers, participants guessed the answer to four questions: (1) “What percentage of votes did Hillary Clinton win in Maryland?” (2) “What percentage of votes did Donald Trump win in New Jersey?” (3) “What percentage of votes did Hillary Clinton win in Alabama?” and (4) “What percentage of votes did Donald Trump win in Utah?” They then indicated their age, sex, and political orientation. The experimenter checked the answers against a sheet with the correct answers to maintain the cover story, but everyone received the coupon for participating.

Results

Six participants did not choose any advice. Participants who chose advice overwhelmingly preferred advice from a diverse group of two people (77%) rather than from a homogeneous group of 3, χ^2 (1, N = 140) = 121.99, p < .001. Political orientation did not predict whether people chose (vs. did not choose) diverse advice. However, of those who chose homogeneous advice, their own political orientation strongly predicted the type of homogeneous crowd that they chose, r = -.59, N = 37, p < .001. The more liberal participants were, the less likely they were to choose conservative homogeneous crowds.

Discussion

Using a paradigm in which people normatively should not prefer diverse advice if they are seeking the most accurate advice, I nevertheless found that most people choose diverse advice. This suggests that people are willing to act on their beliefs that diverse crowds are wiser. This also suggests that they place some measure of value on the accuracy gains that diverse crowds supposedly bring. However, all of the advice in this study was free, so these results do not reveal whether people are willing to invest more resources to obtain diverse advice as opposed to homogeneous advice.

Study IV.4

In many conditions, obtaining information can cost resources including time, energy, and money. People tend to work with similar others (Feld, 1982), so obtaining a diverse crowd can cost more than obtaining a homogeneous one. Although people choose diverse crowds when they cost no more than homogeneous ones, are they willing to choose them when they cost more (H4)? Study 4 tests whether people are willing to pay more for diverse advice when engaging in a political judgment task. In the study, people earn $1 for correct answers and can pay different amounts of money for homogeneous and diverse advice. Participants also report their beliefs about the accuracy of diverse and homogeneous advice.

Method

Participants were recruited from an introductory psychology course. One hundred and five participants took the study for course credit. Their mean age was 18.95 (*S.D.* = 1.14), there were 53 men, 51 women, and 1 participant did not indicate sex. The group was 6% Black, 62% White, non-Hispanic, 5% East Asian, 6% Latino/a, 10% South Asian, Pacific Islander, or other Asian, 2% “Other”, and 10% mixed race. The group leaned liberal, M_{POLIT} = 3.52, *S.D.* = 1.47 (1 = Very liberal, 7 = Very conservative).

Each session included between 1 to 10 participants seated at individual computers. At the start of each session, participants learned that in a previous study I asked 200 American adults to guess what percentage of votes Hillary Clinton and Donald Trump would win in different states in the 2016 presidential election. Participants were told that they would likewise guess what percentage of votes the candidates received and that they could buy different types of advice drawn from that previous study. Each answer that fell within 1 percentage point of the correct answer would earn them $1, and from their bonus they would pay for one randomly chosen piece of advice that they bought. Participants were quizzed on the instructions and proceeded to the task only when they answered all questions correctly.

Twelve survey blocks asked participants to guess the percentage of votes a candidate received in a given state. Block order was randomized. Six questions asked about Trump, and 6 asked about Clinton. Four states had leaned Democrat in the previous election, four had leaned Republican, and two (Ohio and Florida) were traditionally swing states. Four conditions varied which states were paired with which candidates and which type of advice. Across conditions, each state was paired evenly with each candidate, and each type of advice was paired with each state.

In each survey block, participants first read the question and the advice instructions. For example, "What percentage of votes did Donald Trump win in Alabama? For each of the advice pairs below, click on the one you would prefer to buy." They then indicated their overall preference for a homogeneous group's advice vs. a diverse group's advice. There were 4 types of homogeneous advice, crossing sex and political orientation. For every question one type of homogeneous group (e.g., conservative women) was contrasted against a diverse group. For example, "Overall do you prefer to see the average guess from 6 liberal men or a group of 6 men and women, conservatives and liberals?"

For each question participants completed two advice choice sets in random order. One advice set held the cost of homogeneous advice constant at 10 cents while diverse cost varied. The other advice set held the cost of diverse advice constant while homogeneous cost varied. For example, participants indicated their preference for each of these pairs:

(a) liberal men, 10 cents **or** men, women, conservatives, liberals, 10 cents
(b) liberal men, 10 cents **or** men, women, conservatives, liberals, 25 cents

(c) liberal men, 10 cents **or** men, women, conservatives, liberals, 50 cents

(d) liberal men, 10 cents **or** men, women, conservatives, liberals, 75 cents

(e) liberal men, 10 cents **or** men, women, conservatives, liberals, 1 dollar

Participants were told that one of their choices would be randomly presented on the next page. In fact, the same advice was presented to all subjects, with only the label varying according to one of their choices. Thus their perceptions of the effectiveness of advice type could only be derived from the label, not the advice itself. After seeing the advice, participants typed their guess into an open-ended text box.

After completing the 12 blocks, participants indicated how useful they thought each type of advice would be for this type of task. First, they answered, "For the questions like the ones in this in this survey, which of the following group types do you think would offer the most accurate advice overall?" The choices were: "Men and women of all ages," "Groups with specific demographics (like just liberal men, or just conservative women)," or "Both types of groups are equally accurate." (The first item wording was a mistake; it should have read "Men and women, liberals and conservatives" but it nevertheless conveys more demographic diversity than the other option.) The next two questions asked about the frequency with which one type of advice was the best: "For questions like the ones in this survey, how often do you think a group of men and women, liberals and conservatives would offer the most accurate advice?" and a Likert response scale was used (1 = "Almost Never," 2 = "About 25% of the time," 3 = "About 50% of the time," 4 = "About 75% of the time," and 5 = "Almost always.") The same question was asked about specific, homogeneous groups.

The third set of questions measured what accuracy gains or losses people expected for the average person who heeded diverse vs. homogeneous advice. The instructions read: "For the questions below, imagine an average participant named 'Sam'. Suppose Sam always chose and heeded advice from men, women, liberals and conservatives. How do you think this would affect Sam's performance compared to not using any advice?” The options were, “It would make Sam more accurate” (coded as 1), “It would make Sam less accurate” (coded as -1), and “It wouldn’t affect Sam’s accuracy” (coded as 0). A continuous slider scale anchored at 0 and 50 was then presented so that participants could indicate by how much Sam’s score would increase or decrease. Participants who said it wouldn’t be affected were told to indicate 0. In analyses, the answer on the slider scale was multiplied with the answer to the accuracy question to compute an index of accuracy change.

Participants answered the same question for 2 homogeneous groups. For example, “Suppose Sam always chose & heeded advice from (liberal men/conservative women). How do you think this would affect Sam's performance compared to not using any advice?” The group presented for each question was shuffled such that a different group was presented for each question, and the two groups would not overlap on gender or political orientation. Groups were evenly presented across conditions. For analyses, answers from these two questions were averaged together.

At the end of the survey, participants indicated their sex, age, political orientation, and ethnicity. They were paid for any questions they got correct ($M_{PAYMENT}$ = 1.19, *S.D.* = 1.08), and no deductions were taken from their bonus based on their advice choices.

Results

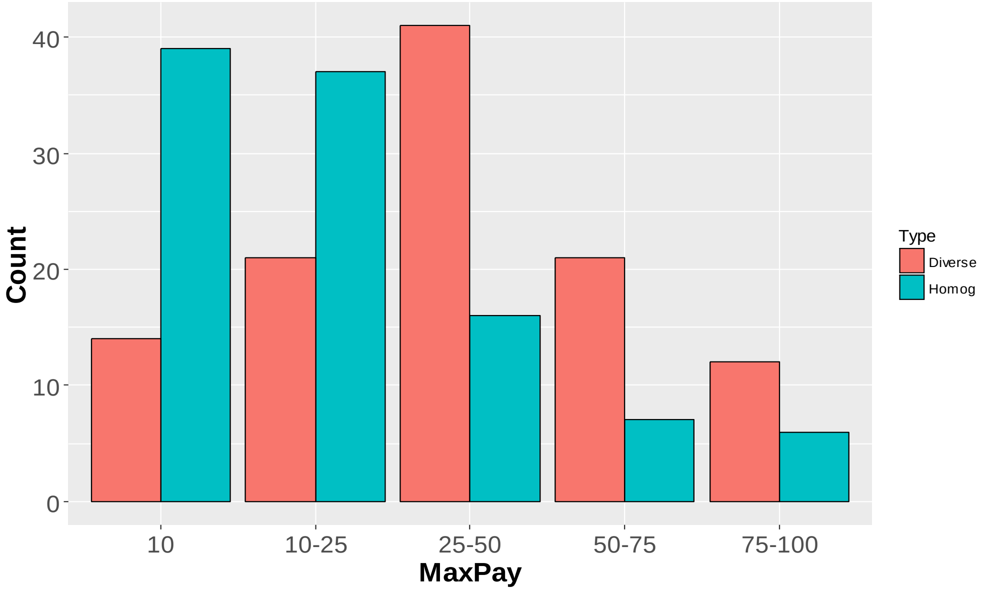
Overall advice preference. Participants strongly preferred advice from diverse sources. Before making their advice choices for each of the 12 questions, participants first indicated what type of advice they preferred overall. The proportion of times that participants chose diverse advice for that question was computed. The results were heavily negatively skewed. The median proportion was .83, and 32% percent of participants always chose the diverse crowd for this question. To analyze the highly skewed data, participants were simply binned into 2 groups – those who chose diverse crowds half of the time or less, and those who chose diverse over half of the time. The vast majority of participants (77%) chose diverse groups over half of the time, χ^2 (1, *N* = 105) = 30.94, *p* < .001.

Equal cost advice. When the cost of homogeneous and diverse advice was equal, people chose diverse crowds most of the time. Recall that for each choice set, either diverse or homogeneous crowds were more expensive, and the other crowd cost was held at 10 cents. Each choice set began with the two crowds costing 10 cents each. The proportion of times that each participant chose the diverse group was computed across all questions. The distribution was again heavily negatively skewed. The median proportion was .75, and 23% of participants always chose the diverse advice for this question. Binning participants into those who chose the diverse group over 50% of the time and those who chose the diverse group 50% of the time or less revealed that 78% of participants chose the diverse crowd over half of the time, χ^2 (1, *N* = 105) = 33.15, *p* < .001.

Maximum willingness to pay. For each of the 12 questions, I computed the maximum participants were willing to pay (WTP) for diverse advice when it cost more and I computed the maximum participants were willing to pay for homogeneous advice when it cost more. This maximum was averaged across all 12 questions separately for diverse advice and for homogeneous advice.

The maximum WTP was not normally distributed. The median WTP for diverse advice when it cost more was 38.33 cents. The median WTP for homogeneous advice when it cost more was 13.75 cents. Thus when directly compared, people are willing to spend 28.33 cents more on diverse advice but only 3.75 cents more on homogeneous advice. Figure IV.4 shows the distributions of maximum WTP for diverse vs. homogeneous advice with the WTP variable binned into 5 categories indicating the maximum amount people were willing to pay for the more expensive advice. For homogeneous advice questions, most people were willing to pay between 10 to 25 cents. People were willing to pay about twice as much for diverse advice: between 25 and 50 cents.

Figure IV.4. Distribution of Maximum WTP by Advice Type, Study IV.4



Note. The *y*-axis indicates the number of participants. The red bars show the distribution of maximum WTP when diverse advice was more expensive, and the blue bars show the distribution when homogeneous advice was more expensive. People had a higher maximum WTP when the crowd was presented as diverse than when it was presented as homogeneous.

A Wilcoxon signed rank test was performed on WTP for diverse vs. homogeneous advice. For each participant, WTP_{DIV} was compared against their average WTP_{HOM}. The difference was significant, $Z = 6.03$, $p < .001$, $r = .64$.

Beliefs about advice. Most participants (50%) thought that for the questions in this survey, diverse crowds would provide the best advice overall. Twenty nine percent thought crowds from a specific demographic group would provide the best advice overall, and 20% thought both types of crowds would be equally accurate. Response proportions significantly differed, χ^2 (1, *N* = 105) = 15.57, *p* < .001. In response to how often each crowd type would be the most accurate, people thought diverse crowds would be the most accurate crowd 60% of the time versus 31% of the time for homogeneous groups. (On the 5-point response scale, M_{DIV} = 3.38, *S.D.* = 1.13, and M_{HOM} = 2.24, *S.D.* = .89). The difference was significant, *t*(103) = 6.97, *p* < .001.

People thought heeding diverse advice would improve accuracy while heeding homogeneous advice would harm accuracy. Specifically, when people considered a hypothetical participant, Sam, who heeded diverse advice, they expected Sam's performance to improve by 10.87 points on average (*S.D.* = 18.12). This improvement was significantly different from 0, *t*(103) = 6.12, *p* < .001. When Sam heeded homogeneous advice, people expected Sam's performance to worsen by 7.06 points on average (*S.D.* = 19.37), and this significantly differed from 0, *t*(103) = 3.72, *p* < .001.

Discussion

Participants were more willing to pay for diverse advice than for homogeneous advice. This difference was consistent with their beliefs that diverse crowds improved performance more often and to a greater degree than homogeneous crowds. In fact, they believed that homogeneous crowds would harm performance. These beliefs are inconsistent with the conclusions from Chapter III – homogeneous crowds help accuracy, and just as much as diverse crowds.

Study IV.5

Study IV.5 tests the same question as Study IV.4 using a similar paradigm, but in a far less adversarial domain. Instead of political judgments, people in Study IV.5 guessed the popularity of different books. Before making their guess for each book, they indicated their preference for advice from diverse versus homogeneous crowds.

Method

Participants were recruited on MTurk to complete a survey on trivia questions. They were told about our previous study in which I asked 50 MTurk participants to rate their interest level for various books. For this study, they guessed the average rating of those 50 participants. Before making each guess, they were presented with several pairs of advice options, and for each pair they indicated which advice type they preferred. Advice varied in cost and whether it was drawn from a homogeneous or diverse crowd. Participants were given a $1 bonus for every answer within .2 points of the correct answer.

Books. Participants guessed the average interest for 8 novels which had been pre-tested as gender neutral in their appeal. The books were: Beyond Black, by Hillary Mantel; City of Bones, by Cassandra Clare; Freedom, by Jonathan Franzen; Harry Potter and the Sorcerer's Stone, by J. K. Rowling; Shalimar the Clown, by Salman Rushdie; The Great Gatsby, by F. Scott Fitzgerald; To Kill a Mockingbird, by Harper Lee; Knife of Dreams, by Robert Jordan. The books were taken from a larger list compiled from various online recommendation lists. They varied in genre as well as publication date, containing both “classics” and modern novels. They also varied in popular appeal, as based on the pre-test.

Advice. Participants were told that their advice was derived from a dataset of approximately 200 people that had also made estimates for these books in a previous study. Ten randomly chosen people’s guesses were averaged together to create each piece of advice, and unbeknownst to them all participants were given the same advice regardless of what option they chose. For each book set, participants were presented with pairs of advice options. They could choose between 1 homogeneous crowd’s advice or 1 diverse crowd’s advice. Four homogeneous crowds were used in this study, crossing age and sex. For a given book, participants could be offered advice from younger women, older women, younger men, or older men. (Four between-subjects versions of the survey were created using a modified latin square design such that across all conditions, each book was evenly paired with each type of homogeneous advice.) Diverse advice was labeled as coming from a group of men and women of all ages.

Procedure. In each survey block, participants first saw the book title, author, and cover image. They then indicated advice preference before receiving advice and making their final guess. For a given book, one type of homogeneous advice was paired with diverse advice across a series of pairs that varied the prices. Participants indicated their preference for each pair.

One set contained pairs in which homogeneous advice cost was held at 1 cent while diverse advice cost varied from 4 to 16 cents in increments of 4. The other set contained pairs in which diverse advice cost was held at 1 cent while homogeneous advice cost varied from 4 to 16 cents in increments of 4[4]. All participants answered both sets of questions for each book, presented in random order. Then, they indicated which advice type they would prefer if both types were equally expensive, at 4 cents. The prices were derived from pre-testing and a previous

[4] For example, participants read, “Choose the option in each row that you would prefer, given the advice type and cost” and then could make a choice for each of the following pairs: (a) younger women, 1 cent or men and women of all ages, 4 cents, (b) younger women, 1 cent or men and women of all ages, 8 cents, (c) younger women, 1 cent or men and women of all ages, 12 cents, and (d) younger women, 1 cent or men and women of all ages, 16 cents. (Vice-versa for homogeneous advice being more expensive.)

study on MTurk in which participants indicated the maximum they would hypothetically be willing to pay for such advice.

After indicating their preferences, participants proceeded to the next page to receive the advice, which they were told was randomly chosen from their preference answers. In fact, they were always given advice coming from whichever group they chose when the price was equal. The advice value was always the same across participants, only the label changed. Participants then gave their answer and proceeded to the next book.

In the final section of the survey, participants indicated how useful they thought homogeneous and diverse advice was for this type of task. The questions were the same as in Study IV.4. They first chose which type of advice would be best overall for this task: (a) advice from men and women of all ages, (b) advice from a specific group, or (c) both types of advice would be equally accurate. They then indicated how often they thought diverse (homogeneous) advice would be the most accurate type of advice on a Likert scale (1 = almost never, 2 = about 25% of the time, 3 = about 50% of the time, 4 = about 75% of the time, 5 = almost always).

Finally, they indicated how different kinds of advice might impact accuracy on the task. The instructions read: “For the questions below, imagine an average participant named ‘Sam’. Suppose Sam always chose and heeded advice from men and women of all ages. How do you think this would affect Sam's performance compared to not using any advice?” The options were, “It would make Sam more accurate” (coded as 1), “It would make Sam less accurate” (coded as -1), and “It wouldn’t affect Sam’s accuracy” (coded as 0). A continuous slider scale anchored at 0 and 50 was then presented so that participants could indicate by how much Sam’s score would increase or decrease. Participants who said it wouldn’t be affected were told to indicate 0. In analyses, the answer on the slider scale was multiplied with the answer to the accuracy question to compute an index of accuracy change.

Participants answered the same question for 2 homogeneous groups. For example, “Suppose Sam always chose & heeded advice from (older men/younger women). How do you think this would affect Sam's performance compared to not using any advice?” The group presented for each question was shuffled such that a different group was presented for each question, and the two groups would not overlap on gender or age. Groups were evenly presented across conditions. For analyses, answers from these two questions were averaged together.

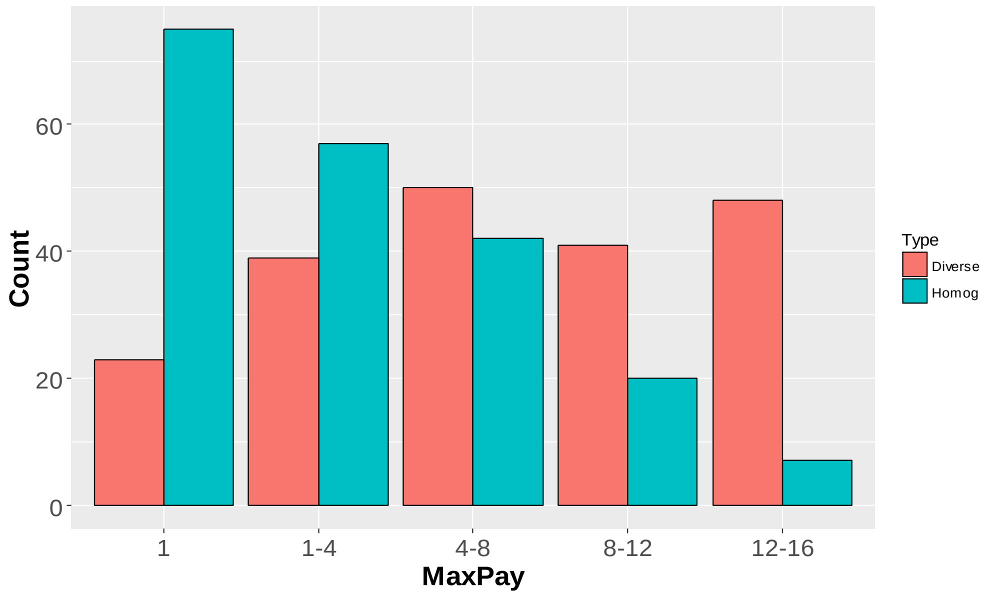
Results

Data collection stopped at 200 participants, which was the intended sample size. However, 25 participants were not directed to the book task but instead immediately proceeded to the final set of questions due to an error in the survey programming. Those participants’ responses were not analyzed and more participants were run to compensate for the error, thus the final sample size included 201 participants who completed the full survey.

Equal cost advice. When the cost for diverse and homogeneous advice was the same, across the 8 books the average proportion of times people chose diverse advice was .75 (*S.D.* = .23). The data were heavily negatively skewed. Therefore, the data were dichotomized into two groups for a simple comparison: participants who chose diverse crowds half of the time or less, and participants who chose diverse crowds over half of the time. The vast majority (78%) chose diverse crowds over half of the time, χ^2 (1, N = 201) = 63.53, p < .001.

Maximum willingness to pay. To analyze WTP, I again computed two variables for each participant: for all 8 books, I separately averaged (1) the maximum WTP for advice when diverse was more expensive and (2) the maximum WTP when homogeneous advice was more expensive. The data were strongly positively skewed for homogeneous advice and diverse advice data resembled a uniform distribution. The median WTP for diverse advice was 7.38 cents (*S.D.* = 4.91), and for homogeneous advice it was 2.25 cents (*S.D.* = 3.48). Figure VI.5 shows the distribution of WTP for diverse and homogeneous advice.

Figure IV.5. Distribution of Maximum WTP by Advice Type, Study IV.5



Note. The *y*-axis indicates the number of participants. The red bars show the distribution of maximum WTP when diverse advice was more expensive, and the blue bars show the distribution when homogeneous advice was more expensive. People had a higher maximum WTP when the crowd was presented as diverse than when it was presented as homogeneous.

A Wilcoxon signed rank test was performed on the data because they violated normality assumptions. Participants were willing to pay significantly more for diverse advice than homogeneous advice, Z = 9.39, p < .001, r = .73. As can be seen in Figure IV.5, when homogeneous advice was more expensive, most participants did not want to pay more for it. The leftmost bin on the *x*-axis, “1”, shows the number of people who always preferred to pay 1 cent for diverse advice as opposed to paying more for homogeneous advice. By contrast, when diverse advice was more expensive, most participants were willing to pay up to at least 8 cents for it.

Beliefs about advice. Most participants (66%) thought that overall, diverse advice was the most accurate. Twenty one percent thought homogeneous advice from a particular group would be the most accurate, and only 13% thought that advice from either group type would be equally accurate, χ^2 (1, *N* = 201) = 96.27, *p* < .001. When indicating how often they thought each type of advice would be the most accurate, participants thought that diverse advice would be the most accurate about 64% of the time (*M* = 3.57, *S.D.* = .90, on the 5-point scale) whereas homogeneous advice would be the most accurate only 36% of the time (*M* = 2.44, *S.D.* = .95). The difference was significantly different, *t*(399.09) = 12.23, *p* < .001.

People thought that the two types of advice would have significantly different effects on performance – that heeding homogeneous advice would hurt the average person's performance, but that heeding diverse advice would improve it. Specifically, participants thought that heeding homogeneous advice would lower Sam's accuracy by 1.46 points (*S.D.* = 3.22), and this decrease was significantly greater than 0, *t*(200) = 6.42, *p* < .001. They thought that heeding diverse advice would improve Sam's accuracy by 2.92 points (*S.D.* = 3.85), and this was significantly greater than 0, *t*(200) = 10.77, *p* < .001. Thus the average difference between expected performance while heeding diverse (vs. homogeneous) advice was 4.38 points, *S.D.* = 5.33.

Discussion

In a different domain with a different sample, this study again found that people act on their beliefs about the superiority of diverse crowds. People thought diverse crowds would be optimal more often than homogeneous crowds, and that heeding diverse advice boosts performance while heeding homogeneous advice harms performance. They were willing to pay more for diverse advice than homogeneous advice although their beliefs about the accuracy gains were wrong.

CHAPTER V

General Discussion

As the model in Chapter II indicates, in order for socially diverse groups to form wiser crowds than homogeneous groups, the groups must be substantially different in the estimates that they provide and the truth must lie close to the midpoint between the means of the two distributions. Across various judgment tasks, I found that these conditions are rarely met. They proved, in fact, to be very difficult to engineer (Chapter III). People of different social backgrounds generally made similar judgments, and the judgment distributions of different groups rarely bracketed the truth neatly. In other words, the diversity (variance) within social groups was relatively big and the diversity (variance) between social groups was relatively small. Homogeneous groups were often as diverse in their estimates as demographically diverse crowds, producing equally wise crowds.

Chapter IV showed that people expect diverse crowds to be wiser than homogeneous ones. Their assumptions about bias seem to be wrong in that they overestimate the effect of social diversity on judgment. When these assumptions are corrected for (Study IV.2), people's inferences lead them to more accurate conclusions. Chapter IV studies also showed that people act on their beliefs about optimal crowds. When using diverse crowds does not cost more, people prefer socially diverse crowds to homogeneous ones (Studies IV.3 – IV.5). When diverse crowds cost more, people were willing to pay about three times as much for diverse advice than for homogeneous advice. Unfortunately, people's beliefs about the relative quality of diverse advice were incorrect, and they consequently invested extra resources on an advice set that did not offer them more accuracy than the alternative.

Importantly, people in Studies IV.4 and IV.5 thought that always heeding diverse advice would *decrease* accuracy. This belief is unsupported by the data in Chapter III. In the original book rating task described in Chapter III, the average individual absolute error was 2.24. Simulated crowd error for homogeneous and diverse crowds were both 1.12. Therefore both types of advice halve error. For the national presidential election task, individual error was 14.52 on average. Diverse crowd error was 9.09 and homogeneous crowd error was 8.89. Again, both types of crowds improve accuracy to about the same degree. Thus, participants in the payment studies underestimated the benefit of homogeneous crowds. One might say that they did not overpay for a diverse crowd, but that they underpaid for a homogeneous one.

What do these findings this say about efforts to increase workplace diversity based on arguments that it is beneficial for performance? The efficacy of such initiatives depends, as I note in the introduction, on the task, what is meant by diversity, and what is meant by performance. For some tasks, demographic diversity can help individual thinking (e.g., Sommers et al., 2008). For our particular task (numerical judgment), and our definitions of performance (accuracy) and diversity (social factors), diversity was not found to be beneficial because the theoretical conditions for its benefits were not met with real data. Thus practitioners must not generalize from the present work to all organizational contexts. They should instead consult empirical work that examines tasks and groups that are analogous to their organization's.

It is also important to note that across all of our tasks and models, social diversity does not *hurt* crowd performance. It was observed to simply have no benefit relative to homogeneous crowds, and theoretically it can only *help* performance. It is also worth noting that, as discussed in Chapter III, people may still prefer diverse crowds over homogeneous ones as a risk reduction strategy. In those presumably rare conditions under which homogeneous groups produce very different estimates, diversity reduces risk by ensuring a moderate performance. People rarely know *a priori* which type of group will be the most accurate (if they do, they of course should poll that group). Thus choosing a homogeneous group carries some risk – it might produce the best crowd, but it just as easily could produce the worst. By contrast, in no simulation or experimental study was the diverse crowd the worst performing one – it was consistently moderate.

Does this work contradict evidence that demographic diversity is beneficial for group performance? As noted in the introduction, social diversity has been found to have measurable benefits in other tasks that are different from those used in the present work. Note, however, that in many of those circumstances, social diversity did not improve performance due to the mechanisms tested here. That is, socially diverse people were not bringing diverse pieces of information to the table. Instead, diversity seemed to motivate the individuals involved to think more critically than they would have in a homogeneous setting. In other studies diversity seemed to lead people to be more open to different perspectives (Phillips et al., 2004). These individual cognitive factors, in turn, could improve group performance.

One limitation of the present work is that none of our tasks involved repeated judgments of the *exact* same phenomenon. Events typically have variance in their realizations – to take the football example, teams' scores vary each year. Thus, to obtain the most rigorous test of the bracketing assumption, I would ideally test predictions of the same game over many years. (I mitigated this problem slightly by using multiple measures in the same domain.) Nevertheless, even if the bracketing condition were to be met on a few occasions for a given event, the social groups in question would still need to be making very different estimates in order for diversity to yield an accuracy benefit. Our many measures suggest that this necessary condition would *not* frequently be met, and its co-occurrence with the bracketing condition would be even less frequent. I am therefore skeptical that a substantial social diversity advantage would emerge over repeated judgments of the same event, but future empirical work is needed to test that proposition.

Another limitation of the present work is that the conclusions rest, in part, on null results, although the more general "wisdom of crowds" effect was consistently observed and there were various significant relationships between social identity and judgment. This does not mean that socially diverse crowds cannot be wiser than socially homogeneous crowds, but I would argue that it is quite difficult to observe and to intentionally create this phenomenon. One may find large differences in judgment if comparing data from rural Vietnam with data collected in downtown Chicago. However, that does not seem to be the magnitude of diversity sought and achieved in many organizations.

Related to the previous point, the findings of null or small effect sizes seem to contradict previous work linking social factors with judgment. People of different cultures think differently (Nisbett, Peng, Choi, & Norenzayan, 2001; Oyserman & Lee, 2008; Savani, Cho, Baik, & Morris, 2015, Yates & de Oliveira, 2016), as do people of different political orientations (Graham, Haidt, & Nosek, 2009; Talhelm et al., 2015). Other ecological factors like disease burden, history of frontier migration, or mode of subsistence have all been linked to social norms that affect thinking, whether directly or indirectly (Gelfand et al., 2011; Talhelm et al., 2014; Uskul, Kitayama, & Nisbett, 2008; Varnum, 2012).

The present findings do not contradict that research. Work finding social differences on judgment has often focused on attitudes or general cognitive style, but the present work focuses specifically on numerical judgments of fact. This is a crucial distinction, as many biases have been shown to be reduced when people are more accountable for their judgments or decisions (e.g., Krizan & Windschitl, 2007; Lerner & Tetlock, 1999). As proposed earlier in this paper, requiring a numerical judgment is a form of higher accountability because the accuracy of the judgment can be verified against a standard. One might therefore expect that most studies of cognition across social groups would find smaller differences for precise judgments about matters of fact than for reports about attitudes, preferences, or “softer” judgment tasks.

Future work may build upon the present findings in several ways. First, the implications of these findings for intergroup dynamics should be explored. How would people respond to the observation that for some tasks, different and even “polar opposite” groups like liberals and conservatives think more similarly than expected? Correcting for people’s over-estimation of differences may promote intergroup harmony and willingness to work with outgroup members. On the other hand, people may take that observation to mean that input from outgroup members is redundant and unnecessary. They may then in turn justify distancing themselves from those people.

Second, the question remains as to whether one can produce the necessary conditions for diversity benefits with intervention or direct manipulation. Processes that amplify biases are good candidates to test. For example, group deliberation about contentious topics in an attitude-diverse group can shift people’s beliefs in more polarized directions (Wojcieszak, 2011). Consequently, judgments about matters of fact might become more biased after people deliberate with outgroup members or come in contact with counterarguments. Such interventions could homogenize estimates within a social group and diversify estimates between social groups, producing one of the conditions necessary for social diversity to outperform homogeneity. It is unclear, however, whether such an intervention would be worthwhile. As individuals become more biased, some or all homogeneous groups may consistently move *away* from good judgment, even as diverse crowds remain fairly accurate due to bias cancellation. In other words, it is possible that polarization would simply make homogeneous crowds worse while leaving diverse crowds unaffected, thereby increasing the discrepancy between those crowd types without necessarily making diverse crowds more accurate.[5]

In conclusion, these findings suggest that while estimate diversity and bias cancellation make crowds wiser than individuals, social diversity does not amplify this wisdom of crowds effect. Thus, social diversity should not be used as a proxy for cognitive diversity in judgment tasks where people are providing concrete judgments independently. In this type of task, when sorted by commonly used social factors like age, sex, and ethnicity, people are more alike than different.

[5] For an illustration, see Figure II.1. At r = .1, assuming the truth is near the middle, participants from both groups are on average fairly accurate. Diverse and homogeneous crowds do well, as do individuals to some degree. At r = .8, by necessity, as groups move away from each other they also move farther away from the truth, again assuming it is close to the middle. Thus, diverse crowds become wiser than homogeneous ones, and individuals fare worse as well. But at r = .8, diverse crowds are ultimately not wiser than either the diverse *or* homogeneous crowds at r = .1.

APPENDIX A

Chapter III Studies: Methods Details

Study 1

Participants were recruited on MTurk to participate in a trivia and prediction task. For this study, I targeted users who reported living in Michigan or Ohio, and analyzed data from participants indicating they cheered either for Ohio State or the University of Michigan's football team (*N* = 51). Respondents were told that the most accurate participant would get a bonus of $5 credited to their MTurk account. They guessed how many points football teams would score in an upcoming game in the 2015 season. The teams and games were presented in random order and included Florida State, University of Florida, University of Connecticut, Temple University, University of Maryland, Rutgers University, University of Wisconsin, University of Minnesota, Northwestern University, University of Illinois, UCLA, University of Southern California, Texas A & M, Louisiana State University, Ohio State University, University of Michigan. Analyses focused on Ohio State and Michigan predictions.

Study 2

Participants estimated the percentage of votes that Republican and Democratic candidates would receive in two upcoming primaries preceding the 2016 presidential election: New Hampshire and Ohio. The most accurate performers (top 10%) were entered into a raffle to win $10. The candidates, presented in random order, were Bernie Sanders, Hillary Clinton, Martin

O’Malley, Donald Trump, Marco Rubio, Ted Cruz, John Kasich, and Jeb Bush. They also indicated how often they read the news. The order in which participants completed the task – demographics or estimation first – was randomized across subjects. The order did not appear to have any systematic effects and it is not discussed further.

Participants also indicated how often they used each of several strategies on a 5-point scale (1 = Never, 5 = Always): Random guessing, web searches, asking another person, reasoning about the question, and going with their gut. People’s dominant strategies included going with their gut (*M* = 3.43, *S.D.* = 1.11) and reasoning about the question (*M* = 3.23, *S.D.* = 1.23). Next, people on average favored random guessing (*M* = 2.34, *S.D.* = 1.23) and reported little use of web searches (*M* = 1.37, *S.D.* = .91) and asking others (*M* = 1.18, *S.D.* = .63).

Study 3

Participants estimated the percentage of Americans that support a variety of polarizing political views. The most accurate participants (top 10%) were entered into a raffle to win $5. They answered six questions: What percentage of Americans (1) “… favor forming a federal database to track gun sales?” (2) “… are against the sale of assault-style weapons to the public?” (3) “… favor building a fence along the entire Mexican border?” (4) “… think the earth is getting warmer mostly because of human activity?” (5) “… favor more offshore oil and gas drilling in the U.S.?” and (6) “... think abortion should be legal in most or all cases?” Respondents typed their answers into an open-ended text box. They also indicated their own attitudes towards each topic; agreement was indicated on a 6-point scale (1 = Strongly Oppose/Disagree, 6 = Strongly Favor/Agree). Finally, participants gave likelihood ratings for each of 7 presidential candidates winning the Iowa caucus preceding the 2016 election. Candidates were Donald Trump, Ted Cruz, Jeb Bush, Ben Carson, Hillary Clinton, Bernie Sanders, and Martin O'Malley.

There were three conditions, with random assignment. All participants completed demographic items first. In one condition, participants reported their personal views before making their estimates for Americans in general. In the second condition, this order was reversed. In the third condition, participants gave their opinions before making estimates and were explicitly reminded of the political orientation that they had indicated. Further, I told them that their views were very important to our study as they might complement the estimates of others such that, when averaged, the group estimates could yield highly accurate results. Accuracy was the same across conditions (p = .734). No condition consistently produced stronger relationships between political identity and estimates. Thus the reported results are collapsed across conditions.

To test for attention, two recall questions asked which items I had (or had not) covered in the survey. Eight percent of subjects failed both questions. They were filtered out, leaving 564 participants.

Study 4

Participants predicted how the candidates would perform in the 2016 United States presidential election. They specifically predicted what percentage of votes Hillary Clinton and Donald Trump would each receive in 10 states in the upcoming 2016 United States presidential election. Four states each had previously favored the Republican (Alabama, Idaho, Utah, Wyoming) or Democratic (Maryland, New York, Vermont, New Jersey) candidate in the 2012 election, and two were typically "swing" states (Ohio and Florida).

Study 5

Participants guessed the popularity rating of 24 books, previously tested among 50 MTurk users. The 24 books were chosen from a larger set of books based on online lists of most popular books. I tested to what extent those books appealed to a particular gender or had gender-neutral appeal. For the judgment task in Study 5, 12 of the pre-tested books had gendered appeal (e.g. “My Sister’s Keeper” is a feminine book), and 12 of the books were neutral (e.g., “Harry Potter and the Sorcerer’s Stone”).

Fifty MTurkers viewed the title, author, and cover image of 24 books and answered the question, “My interest level in this book is…” by choosing an interest level on a scale of 0 to 10, with 1 decimal place. The books varied in how interesting they were to the participants, with average ratings ranging from 2.74 to 6.67. The average interest level in each book was used as the criterion against which the participants in the judgment study were assessed. The participants in the judgment study viewed the title, author, and cover image of the same 24 books and guessed the average interest level for each book on the same scale as the 50 raters.

Study 6

Participants estimated the likelihood of 40 events occurring between the time they took the study (“now”) and May 1, 2016. Questions covered diverse topics such as economics, politics, movies, sports, and social issues, and were derived from current BBC news stories with inspiration from the Good Judgment Project (Mellers et al., 2014). About half pertained to events in the United States and half pertained to international events. The questions were precisely phrased in order to make the outcome as verifiable as possible (see Tetlock & Gardner, 2015 for a good discussion on question specificity and verifiability). Sample questions include: “What is the likelihood of Myanmar's ruling government signing a peace agreement with rebel groups?" "What is the likelihood of California legalizing the recreational use of marijuana?" "What is the likelihood of Leonardo DiCaprio winning an Oscar?" "What is the likelihood of Tesla unveiling the Model 3?" and "What is the likelihood of the value of Logitech stock dipping below $12 USD?"

Participants answered demographic questions first and then completed the prediction task by typing in their estimates into an open-ended text box. They also indicated how often they read the news. Finally, they indicated how often they used each of several strategies on a 5-point scale (1 = Never, 5 = Always): Random guessing, web searches, asking another person, reasoning about the question, and going with their gut. Participants also reported how hard they tried during the task on a 4-point scale (1 = Didn't try at all, 4 = Tried very hard). They were asked to be honest because their answer would not affect compensation.

Most subjects reported trying fairly hard (45%) or very hard (35%) to answer the questions. The most-favored strategies on a scale of 1 to 5 were reasoning about the question (*M* = 3.66, *S.D.* = 1.06) and going with their "gut instinct" (*M* = 3.52, *S.D.* = 1.05). The next most-favored strategy was random guessing (*M* = 2.33, *S.D.* = .99), followed by using information from web-searches (*M* = 1.36, *S.D.* = .82) and asking someone (*M* = 1.19, *S.D.* = .60).

The events were perceived to have varying likelihoods. The event perceived as most likely was the probability of a mass shooting in the U.S. (*M* = 70.05, *S.D.* = 30.48) and the least likely event was Portland, Oregon, being completely submerged in water after an offshore earthquake (*M* = 11.79, *S.D.* = 20.66).

Study 7

Participants of diverse national backgrounds completed three judgment tasks: predicting stock prices, predicting Olympic performance, and predicting news events outcomes. Unlike the other studies which were performed on MTurk, for this study students, faculty, and staff were recruited via targeted e-mails from the registrar's office at a large Midwestern university. Populations of Caucasian Americans, Latin Americans, and East Asians were targeted. Participants who took the study were encouraged to share the study link with their friends, and the study was incentivized with a random prize raffle ($50 Amazon gift card) and an accuracy prize ($100 Amazon gift card). The target *N* was 200 people; 202 people completed the survey (222 including partially completed surveys).

One task involved predicting future stock prices because previous work has found cultural differences in how people predict future trends; Chinese people tend to expect more future change than North Americans and they make different decisions about buying and selling stocks based on those judgments (*14*). Fifteen line graphs showing stock price trends were presented in random order. Each graph plotted prices for November, January, and February. The rest of the months were blank until June. Participants clicked on each image to indicate where they thought the value would be on June 19, 2016.

In the second task, participants then predicted how many medals different countries would win in 9 different Olympic events – for example, "In the 2016 Olympics, how many medals will Brazil win in Judo?" I also asked them how people from that particular country would respond ("What do you think will be the average guess of Brazilian respondents?"). Three questions were about Latin American countries, three were about East Asian countries, and three were about the U.S. teams. All questions paired countries with events for which they had won at least one medal in the previous Olympics.

Finally, participants gave likelihood estimates for 24 events similar to those in Study 6. The events varied in topic and were inspired by news from Latin America, North America, and East Asia. Afterwards, respondents reported their effort (*M* = 2.53, *S.D.* = .61, 1 = Didn't try at all, 4 = Tried very hard). They also reported how often they used each of several judgment strategies as in Study 6: The most popular strategies, in decreasing order, were reasoning about the question (*M* = 3.43, *S.D.* = 1.08), "going with your gut" (*M* = 3.3, *S.D.* = 1), random guessing (*M* = 2.56, *S.D.* = .96), searching the web for information (*M* = 1.45, *S.D.* = .74), and asking someone (*M* = 1.06, *S.D.* = .29).

APPENDIX B

Chapter III Significant Correlations

In the table below, codes for demographic variables are described in the "demographic variable" column. In Study 7, the "Culture" correlations include U.S.-born Americans and Asians born in East Asia. Nationality refers to those born in the U.S. versus born abroad.

Table B.1. Correlations Between Social Identity and Judgment, $p < .01$.

Study	Demographic variable	Question summary	*r*
1	Team OSU fan = 1, UM fan = 0	How many points OSU will score	.379
2	Sex, Male = 0, Female = 1	% of votes Ted Cruz will win in NH primary	.188
2	Education, 1 = Some HS, 6 = Post Graduate	% of votes Bernie Sanders will win in NH primary	.202
2	Polit. Party, 1 = R/ Lib, 2 = Dem	% of votes Donald Trump will win in NH primary	-.205
2	Religious, 0 = non, 1 = religious	% of votes Hillary Clinton will win in NH primary	.198
2	Religious	% of votes Bernie Sanders will win in NH primary	-.215
2	Polit. Party	% of votes Donald Trump will win in OH primary	-.199
3	Political orientation, 1 = Very Lib, 7 = Very Cons	% of Americans favor building a fence along the entire Mexican border	.238
3	Education	% of Americans favor building a fence along the entire Mexican border	-.132
3	Religious	% of Americans favor building a fence along the entire Mexican border	.152
3	Religious	% of Americans favor forming a federal database to track gun sales	.121
3	Age	% of Americans against the sale of assault-style weapons to the public	.143
3	Age	% of Americans favor more offshore oil and gas drilling in the U.S.	.172
3	Political orientation	Likelihood of Trump winning Iowa caucus	.127
3	Political orientation	Likelihood of Carson winning Iowa caucus	.175
3	Education	Likelihood of Clinton winning Iowa caucus	.159
3	Education	Likelihood of O'Malley winning Iowa caucus	-.124
3	Age	Likelihood of Clinton winning Iowa caucus	.188
3	Age	Likelihood of O'Malley winning Iowa caucus	-.187
4	Political orientation	% of votes Clinton will win in Idaho	-.187
4	Political orientation	% of votes Clinton will win in Wyoming	-.180
4	Political orientation	% of votes Clinton will win in Maryland	-.294
4	Political orientation	% of votes Clinton will win in New York	-.383
4	Political orientation	% of votes Clinton will win in Vermont	-.290
4	Political orientation	% of votes Clinton will win in New Jersey	-.294
4	Political orientation	% of votes Clinton will win in Ohio	-.223
4	Political orientation	% of votes Clinton will win in Florida	-.200
4	Religious	% of votes Clinton will win in New Jersey	-.223
4	Political orientation	% of votes Trump will win in Maryland	.251
4	Political orientation	% of votes Trump will win in New York	.401
4	Political orientation	% of votes Trump will win in Vermont	.309
4	Political orientation	% of votes Trump will win in New Jersey	.280
4	Political orientation	% of votes Trump will win in Ohio	.301
4	Political orientation	% of votes Trump will win in Florida	.185
4	Sex	% of votes Trump will win in New Jersey	-.208

4	Religious	% of votes Trump will win in New York	.200
4	Religious	% of votes Trump will win in New Jersey	.195
5	Sex	Guessed interest rating for book: All Quiet on the Western Front	-.204
5	Sex	Guessed interest rating for book: Master and Commander	-.252
5	Sex	Guessed interest rating for book: My Sister's Keeper	.228
5	Sex	Guessed interest rating for book: The Notebook	.227
5	Age	Guessed interest rating for book: Legends of the Fall	.190
5	Age	Guessed interest rating for book: Master and Commander	.266
5	Ethnicity, White = 0, non-White = 1	Guessed interest rating for book: Arthas	.231
5	Ethnicity	Guessed interest rating for book: Pride and Prejudice	.183
6	Age	Likelihood estimates for Germany's parliament restricting migrant privileges	.228
6	Education	Likelihood Raul Castro vacate office	-.272
6	Education	Likelihood Kanye West/Kim Kardashian publicly announce separation/divorce	-.289
6	Education	Likelihood commercial civilian aircraft shot down during armed conflict	-.223
6	Political orientation	Likelihood North Korea has a plutonium bomb	.183
6	Political orientation	Likelihood congress pass stricter gun control laws	.195
6	Political orientation	Likelihood Malawi pass more permissive abortion legislation	.201
6	Religious	Likelihood congress pass stricter gun laws	.183
6	Age	Likelihood Leonardo DiCaprio wins Oscar	.183
6	Sex	Likelihood value of Canadian dollar rise	.186
6	Ethnicity	Likelihood Portland submerged after earthquake	.213
6	Education	Likelihood of more Ebola cases in U.S.	-.192
6	Education	Likelihood lethal confrontation in East China Sea	-.217
6	Education	Likelihood BP stock passing $34 per share	-.184
7	Culture, East Asian = 1, American = 0	Stock value predictions for Office Depot	-.216
7	Religious	Stock predictions for Facebook	.177
7	Culture	Likelihood another Ebola case being reported in US	.296
7	Age	Likelihood another Zika case in South Korea	.217
7	Culture	Likelihood another Zika case in South Korea	-.246
7	Nationality, American = 0, non-American = 1	Likelihood another Zika case in South Korea	-.210
7	Nationality	Likelihood someone surpass Bill Gates as world's wealthiest person	.197

BIBLIOGRAPHY

Antonio, A. L., Chang, M. J., Hakuta, K., Kenny, D. A., Levin, S., & Milem, J. F. (2004). Effects of racial diversity on complex thinking in college students. *Psychological Science*, *15*(8), 507–510. http://doi.org/10.1111/j.0956-7976.2004.00710.x

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership, and men: Research in human relations* (pp. 177–190). Pittsburgh: Carnegie Press. http://doi.org/10.1017/CBO9781107415324.004

Cunningham, G. B. (2009). The moderating effect of diversity strategy on the relationship between racial diversity and organizational performance. *Journal of Applied Social Psychology*, *39*(6), 1445–1460.

Davis-Stober, C. P., Budescu, D. V., Broomell, S. B., & Dana, J. (2015). The composition of optimally wise crowds. *Decision Analysis*, *12*(3), 130–143. http://doi.org/10.1287/deca.2015.0315

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62–68. http://doi.org/10.1037/0022-3514.82.1.62

Eagly, A. H. (2016). When passionate advocates meet research on diversity, does the honest broker stand a chance? *Journal of Social Issues*, *72*(1), 199–222. http://doi.org/10.1111/josi.12163

Eagly, A. H., & Wood, W. (1991). Explaining sex differences in social behavior: A meta-analytic perspective. *qPersonality and Social Psychology Bulletin*, *17*(3), 306–315. http://doi.org/10.1177/0146167291173011

Ellsworth, P. C. (1989). Are twelve heads better than one? *Law and Contemporary Problems*, *52*(1972), 205–224.

Feld, S. L. (1982). Social Structural Determinants of Similarity among Associates. *American Sociological Review*, *47*(6), 797–801. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19586162

Galinsky, A. D., Todd, A. R., Homan, A. C., Phillips, K. W., Apfelbaum, E. P., Sasaki, S. J., … Maddux, W. W. (2015). Maximizing the gains and minimizing the pains of diversity: A policy perspective. *Perspectives on Psychological Science*, *10*(6), 742–8. http://doi.org/10.1177/1745691615598513

Gigone, D., & Hastie, R. (1993). The common knowledge effect: Information sharing and group judgment. *Journal of Personality and Social Psychology*, *65*(5), 959–974. http://doi.org/10.1037/0022-3514.65.5.959

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. http://doi.org/10.1037/a0015141

Harrison, D., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review*, *32*(4), 1199–1228. http://doi.org/10.5465/AMR.2007.26586096

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555–588. http://doi.org/10.1037/a0015701

Heath, C., & Gonzalez, R. (1995). Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organizational Behavior and Human Decision Processes*. http://doi.org/10.1006/obhd.1995.1024

Herzog, S. M., & Hertwig, R. (2014). Think twice and then: combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology*, *40*(1), 218–32. http://doi.org/10.1037/a0034054

Homan, A. C., van Knippenberg, D., Van Kleef, G. A., & De Dreu, C. W. (2007). Bridging faultlines by valuing diversity: Diversity beliefs, information elaboration, and performance in diverse work groups. *Journal Of Applied Psychology, 92*(5), 1189–1199. http://doi.org/10.1037/0021-9010.92.5.1189

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385–16389. http://doi.org/10.1073/pnas.0403723101

Hung, A. A., & Plott, C. R. (2001). Information cascades : Replication and an extension to majority rule and conformity- rewarding institutions. *American Economic Review*, *91*(5), 1508–1520.

Jackson, S. E., Joshi, A., & Erhardt, N. L. (2003). Recent research on team and organizational diversity: SWOT analysis and implications. *Journal of Management*, *29*(6), 801–830. http://doi.org/10.1016/S0149-2063(03)00080-1

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why differences make a difference: A field study of diversity, conflict, and performance in workgroups. *Administrative Science Quarterly*, *44*(4), 741–763. http://doi.org/10.2307/2667054

Ji, L.-J., Zhang, Z., & Guo, T. (2008). To buy or to sell: Cultural differences in stock market decisions based on price trends. *Journal of Behavioral Decision Making, 21*(4), 399–413. http://doi.org/10.1002/bdm

Judd, C. M., Ryan, C. S., & Park, B. (1991). Accuracy in the judgment of in-group and out-group variability. *Journal of Personality and Social Psychology, 61*(3), 366–379. http://doi.org/10.1037/0022-3514.61.3.366

Kahan, D. M. (2010). Culture, cognition, and consent: Who perceives what, and why, in “acquaintance rape” cases. *University of Pennsylvania Law Review*, *158*(29), 729–813. http://doi.org/10.2307/20698345

Kahan, D. M. (2016). The expressive rationality of inaccurate perceptions. *Behavioral and Brain Sciences*.

Kahan, D. M., Hoffman, D. A., & Braman, D. (2009). Whose eyes are you going to believe ? Scott v. Harris and the perils of cognitive illiberalism. *Harvard Law Review*, *122*(3), 837–906.

Kaplan, M. F., & Miller, C. E. (1987). Group decision making and normative versus informational influence: Effects of type of issue and assigned decision rule. *Journal of Personality and Social Psychology*, *53*(2), 306–313. http://doi.org/10.1037/0022-3514.53.2.306

Klein, K. J., & Harrison, D. A. (2007). On the diversity of diversity: Tidy logic, messier realities. *Academy of Management Perspectives*, *21*(4), 26–33. http://doi.org/10.5465/AMP.2007.27895337

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, *67*(4), 596–610. http://doi.org/10.1037/0022-3514.68.4.579

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. http://doi.org/10.1037/0033-2909.108.3.480

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. http://doi.org/10.1287/mnsc.l060.0518

Lau, D. C., & Murnighan, J. K. (1998). Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review*, *23*(2), 325–340.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. http://doi.org/10.1037/0033-2909.125.2.255

Levine, S. S., Apfelbaum, E. P., Bernard, M., Bartelt, V. L., Zajac, E. J., & Stark, D. (2014). Ethnic diversity deflates price bubbles. *Proceedings of the National Academy of Sciences*, *111*(52), 1–6. http://doi.org/10.1073/pnas.1407301111

Lightle, J. P., Kagel, J. H., & Arkes, H. R. (2009). Information exchange in group decision making: The hidden profile problem reconsidered. *Management Science*, *55*(4), 568–581. http://doi.org/10.1287/mnsc.1080.0975

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025. http://doi.org/10.1073/pnas.1008636108

Loyd, D. L., Wang, C. S., Phillips, K. W., & Lount, R. B. (2013). Social Category Diversity Promotes Premeeting Elaboration: The Role of Relationship Focus. *Organization Science*, *24*(3), 757–772. http://doi.org/10.1287/orsc.1120.0761

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. http://doi.org/10.1037/a0036677

Mannix, E. A., & Neale, M. A. (2005). What differences make a difference? The promise and reality of diverse teams in organizations. *Psychological Science in the Public Interest*, *6*(2), 31–55. http://doi.org/10.1111/j.1529-1006.2005.00022.x

McGrath, J. E., Berdahl, J. L., & Arrow, H. (1995). Traits, expectations, culture, and clout: the dynamics of diversity in work groups. In S. E. Jackson & M. N. Ruderman (Eds.), *Diversity in work teams: Research paradigms for a changing workplace* (pp. 17–46). Washington, DC: American Psychological Association.

Mellers, B. A., Ungar, L. H., Baron, J., Ramos, J., Gurcay, B., Fincher, K., … Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–15. http://doi.org/10.1177/0956797614524255

Northcraft, G. B., Polzer, J. T., Neale, M. A., & Kramer, R. M. (1995). Diversity, social identity, and performance: Emergent social dynamics in cross-functional teams. In S. E. Jackson & M. N. Ruderman (Eds.), *Diversity in work teams: Research paradigms for a changing workplace* (pp. 69–96). Washington, D.C.: American Psychological Association. http://doi.org/http://dx.doi.org/10.1037/10189-003

Page, S. E. (2007). Making the difference: Applying a logic of diversity. *Academy of Management Perspectives*, *21*(4), 6–20. http://doi.org/10.5465/AMP.2007.27895335

Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton: Princeton University Press.

Pelled, L. H. (1996). Demographic diversity, conflict, and work group outcomes: An intervening process theory. *Organization Science*, *7*(6), 615–631.

Pelled, L. H., Eisenhardt, K. M., & Xin, K. R. (1999). Exploring the black box: An analysis of work group diversity, conflict, and performance. *Administrative Science Quarterly*, *44*(1), 1–28. http://doi.org/10.2307/2667029

Phillips, K. W., & Loyd, D. L. (2006). When surface and deep-level diversity collide: The effects on dissenting group members. *Organizational Behavior and Human Decision Processes*, *99*, 143–160. http://doi.org/10.1016/j.obhdp.2005.12.001

Phillips, K. W., Mannix, E. A., Neale, M. A., & Gruenfeld, D. H. (2004). Diverse groups and information sharing: The effects of congruent ties. *Journal of Experimental Social Psychology*, *40*, 497–510. http://doi.org/10.1016/j.jesp.2003.10.003

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363. http://doi.org/10.1037/1089-2680.7.4.331

Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301. http://doi.org/10.1016/0022-1031(77)90049-X

Simmons, J. P., & Massey, C. (2012). Is optimism real? *Journal of Experimental Psychology: General*, *141*(4), 630–634. http://doi.org/10.1037/a0027405

Sommers, S. R. (2006). On racial diversity and group decision making: Identifying multiple effects of racial composition on jury deliberations. *Journal of Personality and Social Psychology*, *90*(4), 597–612. http://doi.org/10.1037/0022-3514.90.4.597

Sommers, S. R., Warp, L. S., & Mahoney, C. C. (2008). Cognitive effects of racial diversity: White individuals’ information processing in heterogeneous groups. *Journal of Experimental Social Psychology*, *44*(4), 1129–1136. http://doi.org/10.1016/j.jesp.2008.01.003

Stasser, G., & Stewart, D. (1992). Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment. *Journal of Personality and Social Psychology*, *63*(3), 426–434. http://doi.org/10.1037/0022-3514.63.3.426

Sunstein, C. R., & Hastie, R. (2008). *Four failures of deliberating groups* (Coase-Sandor Working Paper Series in Law and Economics No. 401). Chicago. Retrieved from http://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1213&context=law_and_economics

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

Surowiecki, J. (2005). *The wisdom of crowds*. New York: Anchor Books.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. New York: Crown Publishers.

van Knippenberg, D., & Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology*, *58*, 515–541. http://doi.org/10.1146/annurev.psych.58.110405.085546

Webber, S. S., & Donahue, L. M. (2001). Impact of highly and less jobb-related diversity on work group cohesion and performance: a meta-analysis. *Journal of Management*, *27*, 141–162. http://doi.org/10.1016/S0149-2063(00)00093-3

Windschitl, P. D., Smith, A. R., Rose, J. P., & Krizan, Z. (2010). The desirability bias in predictions: Going optimistic without leaving realism. *Organizational Behavior and Human Decision Processes*, *111*(1), 33–47. http://doi.org/10.1016/j.obhdp.2009.08.003

Wojcieszak, M. (2011). Deliberation and attitude polarization. *Journal of Communication*, *61*(4), 596–617. http://doi.org/10.1111/j.1460-2466.2011.01568.x

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. http://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120. http://doi.org/10.1016/j.obhdp.2006.05.006

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