6-4-2014

Wisdom of Crowds: Tests of the Theory of Collective Accuracy

Scott Ryan
University of Connecticut - Storrs, scott.ryan@uconn.edu

Wisdom of Crowds: Tests of the Theory of Collective Accuracy

Scott Ryan

University of Connecticut, 2014

Organizations have a unique ability to draw on a large collective of individuals to make decisions, judgments, and solve problems. With the complexity of work increasing, a great deal of interest has developed regarding collective leadership in organizations. The term "wisdom of crowds" has surfaced to describe accuracy that can emerge from a large collective of individuals. Collective judgments have been hypothesized to be accurate even when many members of a collective have little knowledge relevant to a judgment. The two most cited predictors of collective accuracy are independence and diversity of judgments. The current project tested three main hypotheses regarding collective accuracy. These hypotheses state that a collective will make judgments that approach zero error, that collective judgments are more accurate when the judgments are made independently than when the judgments are not made independently, and that collective judgments are more accurate when those judgments exhibit diversity than when the judgments do not exhibit diversity. Two experiments involving 33 naturalistic judgments refuted all three hypotheses. Judgments did not approach zero error, collectives composed of independent judges were significantly less accurate than collectives composed of dependent judges, and collectives providing high diversity judgments were significantly less accurate than collectives providing low diversity judgments. The inconsistencies between the current and prior research are explained in terms of the narrow range of judgments used in prior studies, and the lack of specificity in the operational definitions of independence and diversity.

Wisdom of Crowds: Tests of the Theory of Collective Accuracy

Scott Ryan

M.S., Psychology, Brown University, 1999

M.S., Computer Science, University of Rhode Island, 2007

M.A., Industrial and Organizational Psychology, University of Connecticut, 2010

A Dissertation

Submitted in Partial Fulfillment of the

Requirement for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2014

APPROVAL PAGE

Doctor of Philosophy Dissertation

Wisdom of Crowds: Tests of the Theory of Collective Accuracy

Presented by

Scott Ryan

Major Advisor_____

R. James Holzworth

Associate Advisor_____

Janet Barnes-Farrell

Associate Advisor_____

Lutz Hamel

University of Connecticut

2014

Acknowledgments

It is rare to find a mix of both great intelligence and great humility in academic scholars, but I was very lucky to find three individuals who possessed both of these important traits to serve on my committee. First I would like to thank my advisor Jim Holzworth for all of his extensive help over my years of study. From the beginning he has been very supportive and open minded, selflessly allowing me to pursue interests that have brought me great happiness and intellectual growth. I would also like to thank Janet Barnes-Farrell. She has been extremely supportive, open minded, knowledgeable, and quick to respond to any requests for help. I would also like to thank Lutz Hamel, who freely donated his time and wisdom to help on this project. He is a great empirical and theoretical scientist. Dev Dalal, Robert Henning, Vicki Magley, and Steven Mellor have all made my time here a great pleasure, creating a supportive scholarly environment. I would also like to thank my fellow graduate students for all of their intellectual and emotional support. I would not have been able to accomplish my goals without the help of so many great people.

Table of Contents

Wisdom of Crowds: Tests of the Theory of Collective Accuracy

How individuals make collective and individual judgments is an important topic in industrial and organizational psychology (Dalal et al., 2010). Although leaders may have the authority to make judgments on their own, it may be more accurate to encompass the input of subordinates. Leadership research has traditionally focused on characteristics of individual leaders and live interactions with a small number of followers (Yammarino, Salas, Serban, Shirreffs, & Shuffler, 2012). With the complexity of work increasing, there has been an emerging interest in collective leadership (Yammarino & Dansereau, 2008). Collective leadership approaches involve aggregation of skills and knowledge from multiple individuals (Yammarino et al., 2012). One of the best ways to make judgments is to rely on a large collective (Gigone & Hastie, 1997; Larrick & Soll, 2006; Lorge, Fox, Davitz, & Brenner, 1958). A large group of subordinates may have knowledge that supervisors do not possess on their own, and this knowledge may be relevant when making a judgment (Hayek, 1945; Vroom, 2000). The idea of including the input of numerous individuals may have emerged due to a general decline in elitism and distrust of the opinions of experts, who have been shown to have limited accuracy (Meehl, 1954; Shanteau & Stewart, 1992; Tetlock, 2005). With the advent of the Internet and easy access to information, individuals are relying less on the knowledge and opinions of experts.

Although there are advantages to relying on input from others, some leaders may have been put in leadership positions specifically because their knowledge and abilities are greater than those of their subordinates. If leaders believe that their knowledge and judgment ability is superior to their subordinates, it may be wise for them to make judgments independent of input from others (Vroom, 2000). Even if leaders believe that they should include the input of others,

there is still a question of who should be consulted. With modern communication tools, leaders could potentially query thousands of individuals in a matter of hours by posting a survey online or soliciting comments through email. With the concept of collective leadership gaining more research attention (Yammarino et al., 2012), it is important to determine under which conditions collective leadership leads to desirable outcomes.

Until recently, collective judgment studies have primarily focused on groups that interact in a live setting (McGrath, 1984; Stasser & Titus, 2003; Steiner, 1972). With the availability of a new method of aggregation, the Internet, a new interest has emerged in collective judgment. Collective judgment can be contrasted with group judgment in that *group judgment* usually refers to groups that interact in person, whereas *collective judgment* refers to a judgment that uses information from several individuals, whether these individuals interact or not. Much of the interest in collective judgment has been sparked by the influential book *The Wisdom of Crowds* (Surowiecki, 2005). The book has been so influential that the idea that a collective can be more accurate than an individual is often referred to as "the wisdom of crowds." The current project draws on literature from several research areas that investigate collective action in order to formalize the wisdom of crowds into a new theory: *the theory of collective accuracy*. This theory of collective accuracy describes the conditions under which collective judgment is accurate.

Empirical evidence for collective accuracy

The theory of collective accuracy is meant to explain the surprising (Larrick & Soll, 2006) level of accuracy that has emerged from studies of collective judgment. Collective judgment has a long history in social psychology. Much of the research in this area has focused

on the role of interpersonal relationships, interaction patterns, and communication in groups that interact in a live setting (McGrath, 1984; Stasser & Titus, 2003; Steiner, 1972). Some of the earliest work on collective judgment involved groups that did not interact (Gordon, 1924; Stroop, 1932). The idea of using a large group to make simple perceptual judgments, such as size or weight, has a long history in psychology. For example, Galton (1907) studied 787 individuals who entered a contest to guess the weight of an ox. Galton collected the estimates and calculated the median. The median guess was 1,207 pounds, and actual weight was 1,197 pounds. Similarly, Gordon (1924) had individuals rank 10 weights from lightest to heaviest. The correlation between the ranking of the weights and the true values of the weights was computed for each individual. The average correlation was .41. When 50 different individuals ranked the weights, and then the average position was computed, the correlation between these average rankings and the true value improved to .94.

Stroop (1932) replicated and extended Gordon's (1924) study of ranking weights. Gordon averaged the rankings of 50 different individuals. Stroop's study involved a similar task, but used a within-subjects design. He had four *individuals* make 50 different rankings each, and then averaged these 50 rankings from each of these individuals. The correlations between these averaged rankings and the true rankings were .95, .96, .98 and .98. When individuals made only one rating, the average correlation was only .41. These data showed that a large number of rankings, whether provided by the same or different individuals, will be very accurate. In both the within and between subjects measurements, the values will be accurate for the same reason. Random error will cancel, and the value will approach the true mean. This is similar to asking individuals to answer several questions on a scale in order to get a more reliable measure. These results lead to Hypothesis 1:

Hypothesis 1: A large collective will produce judgments that approach zero error.

Empirical and a priori knowledge

The reaction to collective accuracy research has been polar. Interpreting these results in a negative light, Stroop (1932) wrote "Extreme caution should always be exercised in interpreting data which have been treated statistically that the outcomes of statistical manipulation are not mistaken for experimental results" (p. 562). Other researchers (Lorge et al., 1958) lamented "Not until 1932 were the obvious defects of Knight's so-called "statisticized" technique criticized" (p. 345). In contrast, in more recent times, these results have been hailed as extremely important. Page (2007) writes "Many of the specific examples in which collections of people predict correctly seem almost unbelievable" (p. 177).

Philosophers of knowledge (Moser, 1987) have long contrasted the "statistical" and "experimental" results referred to by Stroop (1932). Knowledge that does not need to be verified by experiment is referred to as *a priori* knowledge. It is mathematical or logical knowledge; tautologies are a type of a priori knowledge. An example of a priori knowledge is the fact that anyone who is over 7 feet tall is also over 6 feet tall. This statement does not need to be verified by measuring individuals' heights. In contrast, knowledge that does require observation to be verified is called empirical knowledge. An example of empirical knowledge is the statement "Everyone on earth is shorter than 8 feet tall." This statement requires verification by measuring everyone on earth to confirm that they are less than eight feet tall.

Stroop's (1932) criticism was that Gordon's (1924) results were a form of a priori knowledge and therefore did not require empirical verification. If this were true it would mean that the results could not have possibly come out any differently, not only in practice but in

theory. It would mean that contrary results would be literally impossible, such as someone being both over 7 feet and under 6 feet tall at the same time. This is not the case. It could have been true that averaging the judgments of several individuals could have resulted in less accuracy. For example, there could have been some systematic bias that made some of the weights appear heavier than they were, and this bias could have led the groups with more individuals to make less accurate, not more accurate, judgments. Stroop (1932) tried to demonstrate the a priori nature of the results by demonstrating that within subject averaging was just as accurate as between subject averaging. However, far from being an a priori fact, other researchers have shown that between subject averaging leads to more accuracy than within subject averaging (Ariely et al., 2000).

Although most researchers agree that collective accuracy is at least partly a consequence of statistical rules (Ariely et al., 2000; Gigone & Hastie, 1997), one could argue that this makes these results more, not less, important. If this result were a basic mathematical fact, it would indicate that it is extremely robust. This is a benefit, not a detriment, of this research. Unfortunately, this accuracy is not a certain mathematical fact. If there is systematic bias in the series of judgments, a judgment will not become more accurate as the size of the collective grows (Gigone & Hastie, 1997; Larrick & Soll, 2006; Lorge et al., 1958). When this technique will lead to accuracy and when it will not is an empirical question. If this increase in accuracy were an obvious, infallible mathematical fact, it should be utilized in every aspect of society; however, Sunstein (2005) states that both public and private institutions do not rely on statistical means, but instead rely on deliberating groups.

Intuitions concerning collective accuracy

The high degree of collective accuracy has been shown to be counterintuitive (Larrick & Soll, 2006). Not surprisingly, individuals tend to think of averaging as creating an answer of average *quality*. Larrick and Soll (2006) suggest that this may stem from the representativeness heuristic (Tversky & Kahneman, 1974). For example, if one is going to have surgery performed by an "average" surgeon, one does not think of that surgeon as the best surgeon. This may be why some individuals do not trust voters, because the "average" American is seen as being of only average intelligence. However, the research of perceptual judgment already reviewed (Gordon, 1924; Stroop, 1932) indicates that *averaging* judgments sometimes *adds* to the quality of the judgment. Two types of research highlight this unusual "average is best" idea. Research (Langlois & Roggman, 1990) has shown that when pictures of faces are combined using a computer, the composite face is more attractive than almost all of the individual faces. The average face is actually not average in attractiveness. A similar situation occurs with many emotional intelligence tests, in which the correct answer is defined as the average of a large sample of test takers, referred to as consensus scoring (Mayer, Salovey, & Caruso, 2004; Mohoric, Taksic, & Duran, 2010; Warwick, Nettelbeck, & Ward, 2010). Again, to be average is to be best.

Theoretical justification for collective accuracy

The accuracy of collective perceptual judgments may seem surprising because it seems very difficult to judge the size of an object, such as an ox, exactly. However, this level of accuracy is a consequence of the Law of Large Numbers. The Law of Large Numbers states that as a sample size increases, the sample mean converges to the population mean (Grinstead &

Snell, 1997; Lindgren, 1993). The question becomes: Is the expected value of a judgment of a large population of guessers the actual mean of the judgment? The answer may appear to be that human perception is not precise enough to make exact estimates; however, human perception does not need to be precise, only unbiased. If some individuals are too high, and others are too low, then all of the errors are random, and cancel. The only case in which a perceptual judgment may be inaccurate is when there is systematic bias (Gigone & Hastie, 1997; Larrick & Soll, 2006; Lorge et al., 1958). For example, systematic bias would occur with visual illusions, in which case the population expected value that is approached is not accurate. It may also happen with certain counterintuitive situations, such as those that may occur in quantum physics or relativity.

It is important to note that the Law of Large Numbers does not guarantee collective accuracy. The Law of Large Numbers states that as a sample of events gets larger, the sample will approach the expected value of a population parameter. In the case of judgments, the expected value of the population parameter is the judgment of all individuals. Unfortunately, the judgment of all individuals may be inaccurate, in which case a collective judgment will be inaccurate. The Law of Large Numbers guarantees convergence on the expected value of a population parameter, not convergence to the accurate value of a judgment. For example, the American population believes that the government spends 27% of its annual budget on foreign aid, but it actually only spends 1% (Brodie, 2012). Therefore, if a very large sample of Americans were surveyed, the Law of Large Numbers guarantees that the sample would approach the population value of 27%, but they would be approaching an inaccurate value. In cases like these the Law of Large Numbers actually guarantees collective *inaccuracy*.

In addition to the Law of Large Numbers, the Condorcet Jury Theorem (Condorcet, 1785) also states that increased group size will lead to more accurate judgments given correct conditions. The Condorcet Jury Theorem considers the case in which individuals are making a decision between two outcomes and the final decision will be made by majority rule. The theorem states that if each individual has a probability of being correct above .50, then the more individuals that vote, the more likely the correct decision will be reached. Another important assumption of this theorem is that each voter is independent, i.e. not influenced by other voters. This assumption is in stark contrast to much of the psychology research in group decision making, in which group decision making is usually studied with groups that interact. The Condorcet Jury Theorem may seem obvious at first. If everyone is likely to be right, it is good to have them all participate in the decision. However, note that this theorem will apply if everyone is only 51% likely to be correct. In this case everyone is almost as likely to be wrong as right, and yet including more individuals makes the group decision more likely to be correct. This theorem also indicates that even if a group were solely composed of experts that had a .90 probability of being correct, adding individuals with a .51 probability of being correct would still lead to an increased chance of making the correct decision. This results in a surprising fact, that adding *less* accurate individuals into a collective can make the collective *more* accurate. This analysis is similar to the previous discussion concerning the Law of Large Numbers. The Condorcet Jury Theorem states that as long as there is no bias that would lead to an incorrect answer, then as the sample size gets larger, the correct answer will be approached.

In the case of simple perceptual tasks that do not involve systematic bias, the answer of how to create the most accurate solution is obvious: Have as many people involved in the judgment as possible. The analysis results in a general rule to maximize collective accuracy: If

there is no systematic bias, but only random error, then simply involve as many people as possible (Gigone & Hastie, 1997; Larrick & Soll, 2006; Lorge et al., 1958).

Although in theory there may be situations in which a large collective should make a judgment, in practice it may be difficult to use a very large group. As has been identified by economists (Clemen & Winkler, 1985), there may be large costs associated with involving several individuals in a judgment. In the cases in which costs are high, the size of the collective should be limited.

Selecting based on expertise

The previous sections have demonstrated that there are several advantages to selecting as many individuals as possible to contribute to a collective judgment. Given a large set of individuals, most of this research suggests that all available individuals should be involved in the judgment. However, there is an alternative procedure. A subset of experts could be chosen from the larger collective, and only the judgments of the experts could be used. It may be an advantage to include only the most knowledgeable individuals because the random error associated with their judgments may be smaller than the error associated with less knowledgeable individuals. A small group of experts may be more accurate than a group of similar size that involves non-experts because the variance of their judgments may be smaller, and therefore closer to the true value of the judgment.

Organizations often must choose between relying on a small amount of people with a large amount of information or a large amount of people with a small amount of information (Sunstein, 2006; Surowiecki, 2005). Leaders are usually placed in charge of others because they have greater knowledge, education, or experience. Leaders could be considered to have more

expertise than their subordinates, and therefore may make more accurate judgments than their subordinates. Including subordinates with inferior judgment ability may (or may not) lower the accuracy of the entire collective. As opposed to selecting the entire collective, selecting only experts from the collective may *appear* to be the most accurate way to make a judgment. However, there are several reasons why selecting experts may not be the most accurate way to make a collective judgment.

The first reason is that experts may not be easy to identify (Shanteau & Stewart, 1992; Taleb, 2012). Even if experts with superior judgments can be identified, one of the key factors is how accurately these individuals can be ranked. If the individuals could not be ranked at all, because their errors are all equal, then there are no experts available to select. If individuals could be ranked with great accuracy, such that we are certain that the first judge is better than the second judge, who is certain to better than the third judge, etc., it is more likely that selecting the top or top few judges would lead to more accuracy than selecting all judges.

Even if individuals could be perfectly ranked on judgment accuracy, collective error can still be reduced by including *less* accurate individuals. If leaders make more accurate judgments than all of their subordinates, incorporating the judgment of anyone else will, by definition, indicate that they are incorporating someone with inferior judgment. This may seem to imply that when leaders have superior judgment they should make the judgment without incorporating judgments from others. However, it is possible that including someone with inferior *individual* judgment could improve *mean* accuracy. This could occur if the leader has a lower error than the second best judge but the second best judge has an error that partially cancels the leader's error.

For example:

	Judgment	True Value	Error
Leader	12	10	2
Second best judge	6	10	4
Mean	**9**	**10**	**1**

In this example the leader is twice as accurate as the second best judge, yet combing their judgment by taking the mean still results in a more accurate judgment. This result occurs because although the individual error moves from 2 to 4, the error of the mean judgment is reduced to 1.

The relationship between individual and collective error is represented in the bias variance decomposition equation. The bias variance decomposition equation describes the relationship between mean squared error (MSE), bias (collective error), and variance. In this case bias is represented by the error of a collective; it is the expected value of a collective estimate minus the true value of a parameter. The relationship between average individual judgment error (MSE), collective judgment variance, and average collective judgment error (Bias) can be represented as:

$$MSE = Variance + Bias^2 \quad (1)$$

This equation indicates that the average squared individual judgment error equals the variance of the collective judgment plus the error of the collective judgment squared. A simple manipulation of this equation (Krogh & Vedelsby, 1995) yields an estimate of collective judgment error (Bias):

$$Bias^2 = MSE - Variance \quad (2)$$

Page (2007) describes this as the diversity prediction theorem, and describes the equation as:

Collective Error = Average Individual Error – Prediction Diversity (3)

where Collective Error = $Bias^2$, Average Individual Judgment Error = MSE, and Prediction Diversity = Collective Judgment Variance.

From the previous example of the leader and the second best judge:

	Judgment	True Value	Error
Leader	12	10	2
Second best judge	6	10	4
Mean	**9**	**10**	**1**

MSE = $(2^2 + 4^2) / 2 =$ **10**, Variance = $((12-9)^2 + (6-9)^2) / 2 =$ **9**, so:

$Bias^2 = 10 - 9$

$Bias^2 = 1$

$Bias = 1$

This indicates that adding inferior judges, defined as those with judgments with greater errors, can actually increase the accuracy of the mean judgment, as long as they increase the variance of the judgments. It is important to note that this will not occur if the errors occur in the same direction, such as:

	Judgment	True Value	Error
Leader	12	10	2
Second best judge	14	10	4
Mean	**13**	**10**	**3**

In this case adding an additional judge increases the error from 2 to 3.

Additionally, if the judgment accuracy of the second best judge is far inferior, even if the errors are on opposite sides of the true value, the mean error can then increase, as in:

	Judgment	True Value	Error
Leader	12	10	2
Second best judge	2	10	8
Mean	**7**	**10**	**3**

Again, the addition of the second judge increases the error from 2 to 3.

A non-mathematical way of describing this cancelling of errors would be to consider skills that are complimentary (Steiner, 1972). Although one member of a team may have a superior ability overall, there may be someone else on the team who has a skill in a different area. In this case the "errors" of the team members may cancel, and even though a member of a team may have inferior ability overall, the addition of the person to the team may increase the ability of the team.

In practical situations individuals are interested in making a judgment before the true value is known. If the true value of the judgment is already known, then there is no need to try to estimate it. The previous discussion uses the simplifying assumption that the true value is known, and individuals can be ranked based on accuracy. However, even before the true value is identified, in many cases individuals can be ranked based on other criteria. Experience, knowledge, or skill on previous judgment tasks could be used to rank individuals. These rankings could be used in place of the rankings of accuracy based on the true value. For example, if asking employees to forecast future sales figures, one cannot simply use the answer with the lowest error to forecast the future value, because the event has not yet occurred. If one is interested in using the judgment of the forecaster with the lowest error, one must use other

criteria, such as past accuracy. This makes the task of selecting the most accurate individuals difficult, because it is likely that any criteria will not identify the most accurate judges perfectly, and therefore rankings of accuracy will be approximate.

It is an empirical question, not an a priori question, of whether errors will cancel when less accurate judges are added to a collective. One of the goals of this project is to determine whether the input of others should be solicited even though their judgments may have greater error. This leads to a series of alternate hypotheses. If errors cancel:

Hypothesis 2A: When ranked from most to least accurate, selecting all individuals will result in the most accurate judgment.

If errors do not cancel:

Alternative Hypothesis 2B: When ranked from most to least accurate, selecting the most accurate individual will result in the most accurate judgment.

If errors only cancel for a small set of the most accurate judges:

Alternative Hypothesis 2C: When ranked from most to least accurate, selecting a subset of the highest ranked individuals will result in the most accurate judgment.

Independence

The previous sections have demonstrated that when there is no systematic bias present in a collective, collective accuracy will increase as the number of individuals increase. Aside from systematic bias, one of the most commonly studied factors affecting collective accuracy is independence (Armstrong, 2001a; Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Page, 2007; Surowiecki, 2005). Researchers have stated that for a collective to be accurate, the judgments

made by the collective must be made independently of others in the collective. Dependence is supposed to be lead to inaccuracy when the judgments are dependent both in terms of statistical dependence (Clemen & Winkler, 1985) and in terms of whether the individuals in the collective are interacting with one another (Janis, 1982; Stoner, 1968).

A great deal of psychological research has focused on interacting groups (Davis, 1992; Esser, 1998; Goodman, 1972; Stasser & Titus, 2003). It is difficult to make any claims about interacting groups without directly comparing these groups to non-interacting groups. The prior section has shown that the mean of individual judgments can be very accurate (Gordon, 1924; Stroop, 1932). Therefore, when comparing interacting and non-interacting groups, the best comparison may be the mean of several individuals vs. the result of interacting groups. Group performance is sometimes evaluated by being compared to the best member of the group (Kerr & Tindale, 2004). However, because the best member of the group is identified after the solution is already known, this does not simulate real-world performance, in which the true value is not known before judgments are made.

The idea that a group might not perform as well as the same number of individuals working separately is called process loss (Steiner, 1972). A great deal of research on process loss has involved brainstorming (Paulus & Dzindolet, 1993; Paulus, Dugosh, Dzindolet, Coskun, & Putman, 2002). There is wide support for the result that brainstorming in non-interacting groups creates more unique solutions than brainstorming in interacting groups (Diehl & Stroebe, 1987; Paulus et al., 2002; Wright & Klumpp, 2004). Explanations for this process loss (Paulus et al., 2002) include evaluation apprehension (Diehl & Stroebe, 1987), social loafing (Latan, Williams, & Harkins, 1979), and loss of time due to interacting with others (Diehl & Stroebe, 1987). Process loss has even occurred with the more basic cognitive processes, such as memory

(Weldon & Bellinger, 1997). Weldon and Bellinger (1997) referred to this as collaborative inhibition, in which an interacting pair of individuals recalls fewer items than the sum of a non-interacting pair. Recent research (Wright & Klumpp, 2004) has shown that collaborative inhibition may be caused by the interference of hearing what the other individual has recalled.

Studies of quantitative judgment tasks sometimes indicated process loss, although interacting and mathematical groups were often found to be equal in accuracy (Fischer, 1981; Gigone & Hastie, 1993; Gustafson, Shukla, Delbecq, & Walster, 1973). Gustafson and colleagues (Gustafson et al., 1973) found that the geometric mean of individual judgments was more accurate than the judgment of an interacting group when estimating likelihood ratios of individuals' gender given their height. A cue learning study indicated that group judgments after discussion were virtually identical in accuracy to the mean of individual judgments before discussion (Gigone & Hastie, 1993). A study of loan officers attempting to predict bankruptcy of real-world companies found similar results, with the mean of individual judgments being almost exactly equal to the judgment of an interacting group of similar size. A study of forecasting grade point averages using subjective probabilities (Fischer, 1981) found that interacting and non-interacting groups were nearly identical in accuracy. Overall, there appears to be little difference in accuracy between mathematical groups and groups using discussion.

Another common criticism of group judgment following discussion is the phenomenon referred to as groupthink (Janis, 1972). Janis described groupthink as "A mode of thinking that people engage in when they are deeply involved in a cohesive ingroup, when the members' strivings for unanimity override their motivation to realistically appraise alternative courses of action" (p. 8-9). Groupthink is a strong form of conformity caused by a lack of critical thinking. Individuals strive to conform so much that they do not bring new information into the

deliberation process. Janis suggested that groupthink was responsible for political failures, such as the failure of the Bay of Pigs invasion of Cuba. Sunstein (2005) reviews the 2004 report of the Senate Select Committee on Intelligence that examined the incorrect conclusion that Iraq had weapons of mass destruction. The report explicitly describes the failure as "groupthink," and states that the CIA failed to consider alternative points of view. The CIA actually had formal procedures to prevent groupthink, such as "devil's advocacy," but these procedures were not followed in this case.

Another result of group interaction is group polarization (Stoner, 1968). Group polarization occurs when a group decision becomes more extreme after discussion than it was before discussion. Sunstein (2006) describes a study in which individuals were separated into groups of individuals who were similar in ideology, either all Democrats or all Republicans. The groups discussed the controversial issues of civil unions for gays and lesbians, global warming treaties, and affirmative action. Results showed that individuals had more extreme positions on the issues after discussion than before discussion. A similar result (Myers, 1975) indicated that moderately pro-feminist women will become more extreme in their endorsement of feminism after group discussion. These groups were also shown to be more homogenous after discussion, with less variability in their beliefs concerning the issues that were discussed. Group polarization has been shown to be a very robust phenomenon (Isenberg, 1986).

Do all of these negative results imply that group decision making is inferior to individual decision making? It is important to note the comparisons made in the studies above. Most of the studies either compared deliberating groups to mathematical groups, or compared groups before and after deliberation. Overall, research has shown that group judgments, whether involving discussion or not, vastly outperform individual judgments (Armstrong, 2001a; Gigone & Hastie,

1993; Gustafson et al., 1973; Hinsz, Tindale, & Vollrath, 1997; Lorge et al., 1958; Paulus et al., 2002). The previous section points to the detriments of *group discussion*, not collective judgment in general.

Although a formal meta-analysis may be needed to reconcile the issue of deliberating vs. mathematical group accuracy, this review does seem to indicate that in the majority of cases reviewed, collectives that involve discussion will be less accurate than those that do not involve discussion. This is especially true in brainstorming and memory studies. In judgment studies, there seems to be only a slight advantage for mathematical groups, with most studies showing no evidence for either technique displaying greater accuracy. A large number of reviews of the literature, in various disciplines, have emphasized the idea that deliberating groups are inferior to groups that do not deliberate. Paulus' review of brainstorming (Paulus et al., 2002) indicates that brainstorming produces fewer ideas when done in a group than when done independently. An influential book on the topic of collective judgment (Surowiecki, 2005), which has popularized the term "The wisdom of crowds", states that one of the preconditions for accuracy is that judgments be made independently. Sunstein's (2006) work on collective judgment has a chapter entitled "The Surprising Failures of Deliberating Groups." Page (2007) proves several theorems indicating that large groups are accurate, and many of these theorems assume independence of group members. In a review of combining forecasts, Armstrong (Armstrong, 2001a) writes "Sometimes forecasts are made in traditional group meetings. This also should be avoided because it does not use information efficiently" (p. 433).

Although research involving group discussion seems to support the hypothesis that independent collective judgments are often more accurate than dependent group judgments, research investigating more structured group interaction supports the opposite conclusion. Unlike

groups that allow free discussion, many collective judgments are made through highly structured processes. Computers allow for more structure to be implemented in group judgment, especially the use of iterative feedback. A popular structured technique is the Delphi technique (Rowe & Wright, 1996). The Delphi technique is an example in which dependence leads to higher, not lower collective accuracy. Although there are several modifications of the technique, the general method is that individuals first make a judgment, then are given anonymous feedback concerning the estimates of other individuals in the group, and then make another judgment. The number of iterations can vary, with several rounds of judgment-feedback-judgment. The idea behind the technique is that because group discussion has been shown to have such detrimental effects (Janis, 1982; Paulus et al., 2002; Sunstein, 2006), limiting interaction, while still allowing information sharing, may result in accurate judgment.

Researchers (Rowe & Wright, 1999) have reviewed the accuracy of Delphi techniques relative to several other methods in the field of forecasting. In contrasting the Delphi technique to estimates that were simply mathematically aggregated with no interaction, five studies found Delphi to be more accurate, five studies found no significant difference, and two studies found Delphi to be less accurate. Comparing Delphi to groups using discussion, Delphi was more accurate in five studies, two studies found no significant difference, and one study found discussion groups more accurate. This review indicated that in general the Delphi technique led to more accuracy than groups using discussion or mathematically aggregated groups.

A technique that emerged from the industrial and organizational psychology literature that allows information sharing without some of the biases of live interaction is the stepladder technique (Rogelberg, Barnes-Farrell, & Lowe, 1992). In the stepladder technique, a group starts with only two members. After these members work together on a problem, others are added to

the group one at a time. An important part of this technique is that each individual is given prior information and time to work on the problem before entering the group. This allows for independent evaluation before biasing can occur from others in the group. The stepladder technique requires individuals to combine independent knowledge while also facilitating information sharing. This method has been found to create more accurate solutions than traditional groups interacting in a live setting (Rogelberg et al., 1992).

Although the Delphi and stepladder techniques have advantages over live discussion, the most common type of structured interaction is a market. A market is defined as the means through which buyers and sellers are brought together to aid in the transfer of goods or services (Reilly & Brown, 2009). The markets that will be considered in the current project are prediction and financial markets, such as the stock market. These markets are all very similar because products and contracts are exchanged using a pricing system. Ideally, much of the information available about the value of an asset is reflected in the price per share of a stock (Reilly & Brown, 2009). Stock markets, commodity markets, futures markets, and information markets all use a similar mechanism to aggregate information: prices.

One of the main tenets of economics is not that all *individuals* are rational, but that collectives are rational. This idea is associated with the efficient market hypothesis (Fama, 1970). The efficient market hypothesis states that a market is efficient if the price of products fully reflect all available information. It is important to note that the hypothesis says nothing about how accurate markets are in general, only that they integrate information in an optimal manner. There are two more formal definitions of the efficient market hypothesis (Fama, 1970). The first is that the return on investment for a security is only a function of its risk, and any other fluctuation is random. For example, a very safe bond might only return 2% per year, but a risky

stock, the price of which varies more than the bond, may return 15%. Another formal definition of the efficient market hypothesis is that future return values for an investment can only be modeled by taking the mean historical return and adding random error. In other words, deviation from the mean return value is simply random. What this means is that no individuals have special information that will allow them to gain higher than average returns. Although it is controversial (Gilson & Kraakman, 2003), there is a great deal of support for the idea that most financial markets are efficient as defined by the efficient market hypothesis (Fama, 1998; Fama, 1970; Gilson & Kraakman, 2003; Jensen, 1978). Markets are widely considered to be the most accurate way to make a judgment, and yet market prices emerge from a collective in which individuals are *interacting* with one another through the trading process. Therefore markets are collectives in which individuals make judgments that are *dependent* on others, and yet they are considered to be the most accurate way to make a judgment (Fama, 1998; Fama, 1970; Gilson & Kraakman, 2003; Jensen, 1978). This accuracy is evidence against the premise that independence is necessary for collective accuracy.

Although it is often stated as a fact that independence is necessary for a collective to be accurate (Armstrong, 2001a; Lorenz et al., 2011; Page, 2007; Surowiecki, 2005), this review indicates that the evidence is actually mixed. In some cases dependence leads to lesser collective accuracy (Janis, 1972; Stoner, 1968), and in other cases dependence leads to greater collective accuracy (Fama, 1970; Rowe & Wright, 1996).

Mathematical arguments regarding independence

The prior sections indicate an intricate relationship between independence and collective accuracy, but the relationship is even more complex than has already been stated. Independence

is considered to have an effect through different mechanisms. Some researchers use the word "dependence" to indicate groups that interact through discussion (Janis, 1982; Stoner, 1968), but others consider dependence in a purely statistical manner (Clemen & Winkler, 1985). Statistical independence is defined in relative terms. Data are not dependent or independent, a datum can only be independent of another datum. Events A and B are independent if the probability of A and B is equal to the probability of A times the probability of B (Kac, 1959), i.e., two events are independent if the occurrence of one event does not change the probability of another.

Statistical dependence can affect collective accuracy in several ways. It is often stated (Lorenz et al., 2011; Surowiecki, 2005) that dependence leads to lower collective accuracy, or even that the only way to assure that larger collectives lead to greater accuracy is for the judgments to be made independently of others in the collective. The general idea behind independence increasing accuracy is that the quantity of information is higher when individuals are independent (Clemen & Winkler, 1985). Information theory describes this as the Shannon entropy of data (Shannon & Weaver, 1949). Dependent data have less information content because they can be characterized by a simple function. The concept of data entropy is used to compress files. For example, if a string of zeros 100 bits long is encountered in a file, it can be compressed from 100 bits into a shorter representation such as "100 zeros." If a similar string has a sequence of zeros and ones that do not form a simple pattern, but rather are randomly ordered, and therefore independent of one another, then the information cannot be compressed. Independent data therefore contain more information because they cannot be summarized and represented by a simpler function. Greater information can lead to greater accuracy. Research involving probability judgments (Ariely et al., 2000) has shown that accuracy increases as the number of judgments increases, but this increase is greater if several individuals make single

judgments compared to one individual making several judgments. This difference may have occurred because more information is available when multiple individuals contribute single judgments than when a single individual contributes multiple judgments.

The idea of information entropy is important in collective judgment. The amount of information entropy, and therefore the independence of the judgments from the collective, determines how much improvement is possible by adding more individuals to the collective (Clemen & Winkler, 1985). For example, if everyone provides exactly the same judgment, the data will be completely dependent, i.e. all information can be predicted simply by looking at the judgment of one individual. Adding individuals to the collective cannot improve a judgment in this case because the information is exactly the same as that already represented by the collective. On the opposite extreme, if everyone who could enter the collective has a different judgment, then it is possible (but not certain) that each new individual entering the collective could improve the judgment.

Many researchers (Lorenz et al., 2011; Surowiecki, 2005) take this result to indicate that for a collective to be accurate, the individuals in the collective must be independent of one another. However, the source that is often cited to support this claim (Clemen & Winkler, 1985) *assumes* that the entire collective is accurate. This source goes on to show that assuming all experts are perfectly accurate on average, more precision will be gained as judges are added to the collective if their errors are independent of one another. What this result indicates is that greater accuracy improvement is *possible* if the members of a collective are independent of one another. *What would actually happen with dependent data, if perfect accuracy is not assumed, is not known.* For example, consider a situation in which the value of pi is being estimated. The greatest dependence may emerge from a group of mathematicians, who all agree that the value is

3.14. If non-mathematicians were then added to the sample, they may make independent judgments, but these judgments would lead to inaccuracy because they do not have the knowledge that the mathematicians possess. Dependence could be increased in this case by having the non-mathematicians consult a textbook and discuss what they think the correct value is. This will lead to more dependence in the collective, but this dependence will lead to more, not less accuracy.

A recent article (Lorenz et al., 2011) suggests that social influence can "undermine" collective accuracy. However, the article only indicates that that *variance* of judgments is decreased in cases of social influence. The reason the authors state that social influence undermines collective accuracy is because collective accuracy is supposed to be associated with higher variance (Armstrong, 2001a; Page, 2007; Surowiecki, 2005). However, their own data do not indicate that this is the case. The results indicated that the accuracy of the collective judgments is actually slightly *higher* in cases of social influence, although the article does not indicate whether this difference is significant. The assumption in this article is that higher variance is a beneficial result, but this is a questionable assumption. This is yet another case in which assumptions take precedence over empirical results, and another case in which is it not clear that dependence is associated with a lack of accuracy.

The issue of dependence is clearly controversial. Whether considering dependence as live interaction or statistical dependence, there appear to be contradictory results in both cases. Several authors (Armstrong, 2001a; Kahneman, 2011; Lorenz et al., 2011; Makridakis, Hogarth, & Gaba, 2010a; Page, 2007; Surowiecki, 2005) plainly state that independence is a necessary condition for a collective to be accurate. However, several other researchers (Fama, 1970; Rowe & Wright, 1996) suggest that sharing information can lead to more accuracy. This conflict could

be framed in terms of what individuals do with information that is acquired from others. Information could be used in a positive way to inform a judgment, or could be used in a negative way to bias a judgment. This leads to two alterative hypotheses. If information sharing biases judgments:

Hypothesis 3A: A collective will be more accurate when judgments are made independently than when made dependently.

If information sharing increases the knowledge of individuals:

Alternative Hypothesis 3B: A collective will be more accurate when judgments are made dependently than when made independently.

The effects of judgment diversity on collective accuracy

The topic of diversity is similar to the topic of independence. Diversity is another cited prerequisite for a collective to be accurate (Page, 2007; Surowiecki, 2005). However, like independence, it is not clear whether the diversity of a collective will be positively or negatively related to the accuracy of a collective. When discussed in the domain of collective accuracy diversity is often defined in terms of the differences between judgments, as measured by the statistical quantity variance (Krogh & Vedelsby, 1995; Page, 2007). Diversity is similar to independence in that both will increase information entropy (Shannon & Weaver, 1949). For example, if judgments were all equal, both the variance and information entropy would be zero. If variance were very high, then everyone would have dissimilar judgments, and information entropy would be high. More information would be available to be incorporated in the judgment, as is the case when judgments are independent.

As previously stated, there is a direct relationship between collective accuracy and variance:

$$\text{Bias}^2 = \text{MSE} - \text{Variance} \quad (2)$$

Page (2007) describes this as the diversity prediction theorem, and describes the equation as:

$$\text{Collective Error} = \text{Average Individual Error} - \text{Prediction Diversity} \quad (3)$$

and makes the statement "Being different is as important as being good" (p. 208). After deriving the same equation, which they call ensemble generalization error, and applying it to results from neural networks, Krogh and Vedelsby (1995) state "We want networks to disagree!" (p. 233).

Reading these sources and observing these equations, it may appear that as the variance of a set of collective judgments increases, the accuracy of the collective judgment increases. However, this will only occur if variance increases *while MSE does not change.* Unfortunately, MSE and variance tend to be positively associated (Meir, 1995). When variance increases, MSE tends to increase. Therefore, increasing variance can increase MSE, resulting in a *less* accurate collective judgment. These results indicate that the ideal situation would be to increase variance of a collective while keeping the average error of all of the individuals constant. Unfortunately increasing variance can increase MSE, and that can decrease collective accuracy (Meir, 1995). These equations therefore cannot illuminate what will happen with real-world collective judgment.

Armstrong (2001) is another researcher who suggests using "heterogeneous" experts when making judgments. Again this will lead to accuracy, but only if certain assumptions are met. If the experts being used are of equal accuracy, then higher variance will guarantee higher

collective accuracy. This is a consequence of bias variance decomposition equation discussed previously. The worst case scenario would be one in which all individuals are equally inaccurate and there is no variance in their judgments. In this case variance would be zero, therefore collective error would equal the average error. This would mean that collective error could not, even in this worst case scenario, exceed the average individual error, assuming all individuals are equally accurate. If the experts were heterogeneous, and their average error were the same, then the variance would not be zero, and therefore the average collective error would be reduced by combining experts. All of this again depends on a questionable assumption, that the experts are equally accurate.

Another example of low vs. high variance estimates involves the differences between within and between subjects estimates; i.e. a set of estimates made by one individual vs. a set of estimates made by several different individuals. Taking a typical collective accuracy example, consider a group of individuals estimating the number of beans in a jelly jar. Due to statistical dependence, one might expect that 100 repeated measurements from the same individual would be less varied than 100 measurements from 100 different individuals. Researchers (Ariely et al., 2000) have shown that between subject estimates of probabilities are more accurate than within subject estimates of probabilities. This could happen because one individual may have consistent bias in a particular direction, which may not cancel itself out. However, several individuals' bias may occur in different directions, which will cancel. Again, the variance of between subjects estimates tends to be higher, which can be associated with greater accuracy.

This analysis has shown that the relationship between the variance and accuracy of a set of collective judgments is an empirical, not an a priori, question. As with all a priori questions, mathematical results will always depend on assumptions that may or may not hold in the real

world. Unfortunately, most of the results discussed with respect to both independence and diversity assume perfect accuracy, although the level of accuracy is the quantity of interest. Most of these mathematical results suggest that higher variance will be associated with greater accuracy. However, there are a number of reasons why *lower* variance would be associated with greater accuracy.

The Central Limit Theorem states that a sample with smaller variance is more likely to accurately estimate the expected value of a population parameter than a sample with a larger variance (Lindgren, 1993). If the population expected value is equal to the accurate value of a judgment (perfect accuracy), then a smaller variance will lead to more precision. However, the expected value of a population parameter may not be an accurate judgment value, so this relationship between low variance and high accuracy will not always hold.

Lower variance may also be associated with greater accuracy because variance may represent the level of difficulty of the judgment. Take again the example estimating the number of jelly beans in a jar. If there are only two beans in the jar, we would expect an extremely low variance from a set of judgments, possibly zero. This low variance would emerge from the fact that the task is so easy that the answer is obvious. If instead there were thousands of beans in the jar, we would expect a much higher level of variance because the judgment would be more difficult.

In summary, there appears to be evidence that high variance of a set of collective judgments can lead to greater accuracy, but other evidence that high variance leads to lower accuracy. We are again faced with alternative hypotheses. If variance increases the amount of knowledge available in a collective:

Hypothesis 4A: The accuracy of a collective will be positively correlated with the variance of the individual judgments.

If variance is associated with reduced precision of a judgment:

Alternative Hypothesis 4B: The accuracy of a collective will be negatively correlated with the variance of the individual judgments.

Forecasting as a judgment task

There are several types of judgment tasks that can be used when researching judgment accuracy. Forecasting future events is a desirable judgment task because it often has a clear criterion and is of high difficulty. Forecasting is a common task in industrial and organizational psychology. Forecasts must be made when selecting or promoting employees, in which future performance must be predicted. Organizations are interested in forecasting several other events, such as the success of various products or the general state of the economy. In a recent review of forecasting results, researchers (Lawrence, Goodwin, O'connor, & Onkal, 2006) suggest that two of the most important areas for future research are the value of expertise and the effects of differences in availability of information. As indicated in this review, these two areas have yielded contradictory results. These are the two primary areas explored in this project.

Experiment 1

Experiment 1 tested several of the hypotheses derived from the theory of collective accuracy:

Hypothesis 1: A large collective will produce judgments that approach zero error.

Hypothesis 2A: When ranked from most to least accurate, selecting all individuals will result in the most accurate judgment.

Alternative Hypothesis 2B: When ranked from most to least accurate, selecting the most accurate individual will result in the most accurate judgment.

Alternative Hypothesis 2C: When ranked from most to least accurate, selecting a subset of the highest ranked individuals will result in the most accurate judgment.

Hypothesis 3A: A collective will be more accurate when judgments are made independently than when made dependently.

Alternative Hypothesis 3B: A collective will be more accurate when judgments are made dependently than when made independently.

Hypothesis 4A: The accuracy of a collective will be positively correlated with the variance of the individual judgments.

Alternative Hypothesis 4B: The accuracy of a collective will be negatively correlated with the variance of the individual judgments.

One of the most controversial questions to emerge from this review concerns independence. The recent trend in large literature reviews (Page, 2007; Sunstein, 2006; Surowiecki, 2005) is to emphasize the importance of independence in collective judgment. Letting individuals make decisions in isolation leads to a lack of bias being introduced from others. This lack of bias is one of the key factors that leads to accurate collective judgment. If individuals make judgments that are not independent, their errors may be similar, and therefore not cancel. However, the accuracy of the Delphi technique (Rowe & Wright, 1999) and markets

(Fama, 1970) may indicate that systematic feedback from others in the collective can lead to learning and increased accuracy.

Experiment 1 tested the effect of independence by providing participants feedback from others in the collective as the experiment progressed. The averages of the previous forecasts were reported to participants as they were making their forecasts.

Method

Participants. Participants were 84 (52% women) undergraduates from a large university in the northeastern U.S. The most common major course of study was psychology with 10% of all participants majoring in psychology. They participated in exchange for partial course credit. Participants were informed that they were not required to participate in the study and had the right to stop participating at any time. The median study completion time was nine minutes. Participants made predictions during the early winter of 2012.

Design. The experiment consisted of two conditions, the control and dependent conditions.

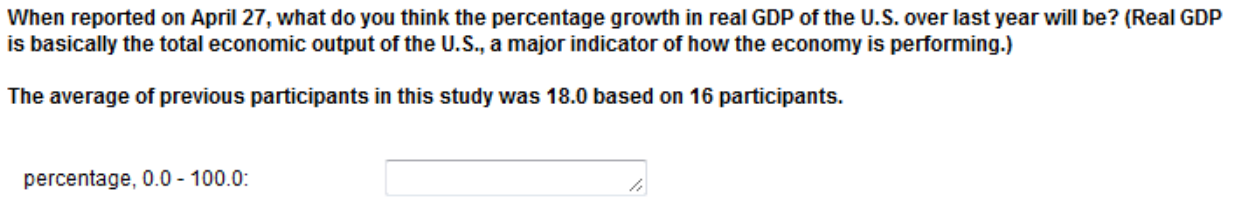
Measures. Measures are described below and displayed at the end of Appendix A.

Demographic and individual difference measures. Participants were asked for their gender, major, and verbal and math SAT scores. Sixteen other items were measured on 5-point Likert scales anchored from strongly disagree to strongly agree. Nine items (three per dimension) were taken from Tetlock (2005) regarding faith in free markets, optimism about the world economy, and the hedgehog-fox dimension (the extent to which an individual uses a single rather than multiple theories to predict events). A single-item happiness measure was taken from

Lykken and Tellegen (1996). Two measures of analytical vs. intuitive thinking styles were included (Holzworth, 2002).

Confidence. To measure confidence in their predictions participants were asked: "What is your percentage of confidence in the previous prediction? Please enter a number from 0 to 100."

Dependence manipulation. Approximately four participants performed the study each day. In the dependent condition, participants were provided with the mean value and sample size from all previous participants for each prediction. This manipulation was meant to represent a dependent judgment in which individuals gather information from others. For example, on the second day, participants in the dependent condition were told that the average from previous participants was 1.11 based on 4 participants, and on the third day, participants were told that the average from previous participants was 2.11 based on 8 participants, etc. For example:



In the control condition participants did not receive information regarding the judgments of other participants. The experiment was run twice with two different samples so that different running means would emerge. The first sample was run for six days and the second sample was run for five days.

Target events. Participants were asked to predict future values of several target events. The events were chosen to include both high and low difficulty judgments with clear numerical

true values. Participants were asked to predict one future value for each of the following quantities (all are for the U.S.A. unless otherwise noted):

- U.S. Gross Domestic Product (GDP) percentage change
- China's GDP percentage change
- Unemployment rate
- Retail sales change
- National average gas price
- Gold price
- Number of homes sold
- Number of unemployment claims
- The movie *The Avengers* money earned
- The movie *Men in Black 3* money earned
- The Dow Jones stock index percentage change over one month
- Apple's stock percentage change over one month
- General Electric's market capitalization
- General Motor's market capitalization

Procedure. See Appendix A for an example of the dependent condition. The study was performed entirely online. After registering for the experiment on a university website, participants were emailed a link and told that they had 24 hours to take the survey. Participants took the survey at the location of their choice. Participants were given a three digit code to enter when they started the survey so that they could be given partial course credit. Participants took the survey using the online survey tool *Qualtrics*©.

At the start of the experiment participants read an information sheet informing them that they had the right to withdraw from the study at any time. Participants were then told to enter the code they were emailed so that they can be given partial course credit. In the control condition participants were told:

> We are going to ask you to make a series of predictions. Please use all of your personal knowledge, intuitions, and reasoning ability to make the predictions. Please do not look up additional information as you are making predictions.

In the dependent condition the following was added:

> You will be provided with the average response from other participants who have already made these same predictions. Please feel free to use that information if you wish. It's up to you.

Next participants were told "When asked for numbers please write only numbers and not symbols like % or $. Feel free to use decimals or not. You can input positive numbers (like 5.0) or negative numbers (like -5.0)."

Participants then made predictions and completed all of the individual difference measures (see Appendix A).

Results

All variables were sorted and examined for outliers. Visual inspection was used to determine if any values were in the far tails of the distributions. Visual inspection determined that values that were 10 standard deviations from the mean appeared to very far from the center of the distribution. Therefore these values were removed before analysis began.

Computation of prediction error measures. For many analyses participants were compared on both collective error and individual error. Individual error was measured by taking the individual judgment, subtracting the true score, and taking the absolute value of the difference. Collective error was measured by first computing the mean (or for some analyses the median) of the entire collective, subtracting by the true value, and then taking the absolute value of the difference. For some analyses absolute percentage error is presented. Absolute percentage error (Armstrong, 2001b) is computed by taking the error as described previously and dividing by the true value and then reporting this value in percentage form.

Collective accuracy. Hypothesis 1 stated that a large collective will produce judgments that approach zero error. Tables 1 and 2 display the true values, mean values, and the error values from samples 1 and 2. The values do not approach zero error. The mean percentage error from all target events combined was 691%. The 95% confidence interval ranged from 224% to 1157%, indicating that mean error was significantly greater than zero. Only 5 of the 64 (16 target events x 2 conditions x 2 samples) judgments had errors lower than 10%. Hypothesis 1 was refuted in this data set.

Selecting accurate subsets. Hypotheses 2A, 2B, and 2C presented three options for selecting accurate subsets of individuals from a collective, assuming that individuals could be ranked based on some (imperfect) predictor of accuracy. These hypotheses stated that a collective would be most accurate if either: (A) all individuals were used, (B) the judgment of the highest ranked individual was used, or (C) some subset was used. In order to perform this analysis, individuals needed be ranked based on accuracy. In real-world tasks the true value is not known when predictions are made, so individuals cannot be ranked based directly on prediction accuracy. For example, in asking 100 individuals to predict U.S. GDP growth for an

upcoming quarter, one cannot simply select the most accurate individual's prediction as a forecast, because the event has not yet occurred and there is no way to know which of the 100 individuals is the most accurate. However, the top individual could be chosen on other predictors of accuracy, such as how well the 100 individuals have performed on similar predictions. In the current study individuals were ranked on how well they performed on similar predictions. Table 3 displays the target event, the similar prediction that was used to rank individuals, and the correlation between the target event prediction accuracy and the similar prediction accuracy. Similar predictions were chosen a priori based on the judgment of the author. If a similar prediction was not significantly correlated with the prediction, other similar predictions were examined until one was found that was significantly correlated with the prediction. Predictions that did not significantly correlate with any other prediction were excluded from the analysis. For example, GM market cap accuracy was not significantly correlated with GE market cap accuracy, so GE Market cap was not used as its similar predictor. However, GM market cap accuracy was correlated with U.S. GDP accuracy, so U.S. GDP was selected as its similar predictor. Once a similar prediction was chosen, individuals were ranked based on how accurate they were on a similar prediction.

Once individuals were ranked based on how accurately they performed on a similar prediction, three sets were selected to test Hypotheses 2A, 2B, and 2C against one another. The three sets were the highest ranked individual, the highest three ranked individuals, and all of the individuals combined. The three highest ranked individuals were chosen as a subset because if all possible subsets were tested one of these would likely be the most accurate simply by chance. Once these three sets were chosen, the error of each group was computed by computing the absolute value of the difference between the mean collective judgment and true value. These

error values are displayed in Table 3. A binomial test indicated that no group was accurate more often than would be expected by chance (33%), $p = .53$. The strongest form of the theory of collective accuracy would suggest that the entire collective will be more accurate than any subgroup in every case. In the current study this would imply that the "all" collective would be the most accurate in 11 out of 11 instances. The "all" collective was the most accurate collective in only 18% (2 of 11) of the cases. A binomial test indicated that this 18% accuracy level was significantly lower than 100%, $p < .001$. Not only was the entire collective the most accurate less than 100% of the time, they were less likely to be accurate than either of the other subsets, although not significantly so. These results indicated that it was possible to select a subset that was more accurate than the entire collective.

Comparing control and dependent conditions. Hypothesis 3A stated that independent judgments would be more accurate than dependent judgments, whereas Hypothesis 3B stated that dependent judgments would be more accurate than independent judgments. Because the distribution of the variables was heavy-tailed, the Mann-Whitney U test was used to compare the medians from the control and dependent conditions. Table 4 displays the median individual judgment errors and the results of significance tests comparing the control and dependent condition errors for each target event. In 2 cases the control condition was significantly more accurate than the dependent condition, but in 12 cases the dependent condition was significantly more accurate than the control condition. Overall, dependence led to more accuracy, but not for every event. These results were not due to inflation of Type I error associated with multiple testing. Only 5% of the tests would be significant assuming chance, but 39% of the tests were significant. The binomial test indicated that this .39 probability of significant tests was greater than the .05 expected by chance, $p < .001$ (Brožek & Tiede, 1952; Hedges & Olkin, 1980).

In order to determine whether the control condition was more accurate than the dependent condition at the collective level, collective accuracy was also examined. Collective accuracy differences between the dependent and control condition were tested by first determining whether median judgments differed in the control and dependent conditions and then comparing which of the medians were closer to the true value (see Table 5). In 3 cases the control condition was significantly more accurate than the dependent condition, but in 9 cases the dependent condition was significantly more accurate than the control condition. As with individual accuracy, overall the dependent condition was more accurate than the control condition. These results were not due to inflation of Type I error. The binomial test indicated that the .33 (12/36) probability of significant tests was greater than the .05 expected by chance, $p < .001$ (Brožek & Tiede, 1952; Hedges & Olkin, 1980).

The effect of time. The mean judgments as a function of time and condition and the correlation between the mean values and time are displayed in Table 6 and Table 7. These means represent the aggregate mean up to that point in time, i.e. the overall mean for that time period including all previous time periods. The absolute value of the correlations between time and the means are high, with 21 of 64 correlations over .90. In both conditions, the means tend to move towards a specific value. This is likely a consequence of the Law of Large Numbers, because the running means are becoming closer to the true population value as the sample size is increasing. This phenomenon represents regression towards the mean. This regression towards the mean even occurred in the dependent condition, where the assumption of independence was violated. This result indicates that as more individuals are added to a sample, the sample will approach the population value even if judgments are not made independently.

Relationship between accuracy and variance. The primary effect of providing the mean judgments of other participants in the dependent condition was to reduce variance of the judgments. Table 8 displays the standard deviations in the dependent condition to the standard deviations in the control condition. The variance was greater in control than in the dependent condition in 26 of the 32 instances, and significantly greater in 13 of the 32 instances. There were no cases in which the dependent condition had a significantly higher variance than the control condition. Because the dependent condition had both greater accuracy and lower standard deviation, these data are evidence for Hypothesis 4B, that lower variance is associated with greater accuracy.

Discussion

Hypothesis 1 stated that a large collective will produce a judgment that is nearly 100% accurate. This hypothesis was refuted. The judgments from this study were difficult and it is likely that participants had little knowledge of their values. However, it is important to note that even if participants had little knowledge of the judgments, the collective could still be accurate. If the errors were random, they would cancel, and the collective judgment would still be accurate. This cancelling of error did not occur in the current study. Individuals tended to be biased by systematically overestimating values in some cases, and underestimating values in other cases. It is true that if individuals guess "randomly" their errors will cancel and a judgment can still be accurate, but when making judgments of continuous unbounded quantities, it is difficult to guess "randomly." When participants are guessing between discrete outcomes, such as a coin flip, it may be easier to simply randomly guess one of two outcomes. In contrast, when guessing a quantity that could take on any value, it may be difficult to guess randomly. Without random guesses, the errors did not cancel, and the collective was not accurate.

Hypothesis 2A stated that a large collective would be more accurate than a subset of the collective, even if those in the subset had lower individual errors than those in the entire collective. This hypothesis was not supported. The most accurate collectives were those composed of either the highest or the three highest ranked individuals. Although it is mathematically possible that adding less accurate individuals into a collective can make the collective more accurate even if these individuals are lower in accuracy, this did not occur in the current study. The results regarding the accuracy of subsets were not conclusive, so a replication will be attempted in Experiment 2.

It could be argued that the reason these judgments were inaccurate is because they were forecasts rather than judgments of current values. Forecasts may be considered to be of higher difficulty. However, most of these judgments changed by only a few percent from the beginning to the end of the study, so simply guessing the current correct value would have led to a very accurate judgment. Inaccuracy in the current study emerged from not correctly judging the current value rather than not correctly judging the future value.

Hypothesis 3A stated that independent judgments would be more accurate than dependent judgments, whereas Hypothesis 3B stated that dependent judgments would be more accurate than independent judgments. For the majority of judgments in the current study the judgments made in the dependent condition were significantly more accurate than the independent judgments made in the control condition. This is likely to have occurred because these were difficult judgments, and many participants did not have any knowledge of the true values. They therefore used the information from others as a cue, and this made them more accurate. The dependent groups also had a lower variance, but were more accurate. This result

supports Hypothesis 4B, that lower variance in a collective is associated with greater collective accuracy.

The fact that there were very high correlations between the aggregate judgments and time in both the dependent and independent conditions indicate that the Law of Large Numbers was valid even under the dependence found in the dependent condition. This may seem to contradict the Law of Large Numbers assumption of independence. However, the independence assumption of the Law of Large Numbers only guarantees validity, it does not necessarily state that the Law of Large Numbers is invalid under dependence (Birkel, 1988). In the current study judgments were approaching the population expected value even under dependence. This result questions the idea that independence of judgments is necessary for collective accuracy, because it appears that even under dependence a large collective will approach the expected value of the population parameter.

The results of this study offer a consistent refutation to the theory of collective accuracy. All four hypotheses suggested by the theory of collective accuracy were refuted. The collective did not display an accuracy approaching zero error, a large collective was not more accurate than a smaller collective ranked on accuracy, independent judgments were not more accurate than dependent judgments, and greater variance was not associated with greater accuracy.

Experiment 2

In Experiment 1 all four of the hypotheses derived from the theory of collective accuracy were refuted. These were not null results; not only was there no evidence for these hypotheses, but in all four cases there was evidence for alternative hypotheses. Independence and diversity have been identified as predictors of collective accuracy (Lorenz et al., 2011; Page, 2007;

Surowiecki, 2005), such that the collective will be more accurate if the individuals in the collective make judgments independently and the judgments are diverse. Experiment 1 indicated that independent judgments were not more accurate than dependent judgments, and diversity, defined as variance, was associated with less accuracy. What instead appeared to be a predictor of collective accuracy was *knowledge*. Many of the judgments in Experiment 1 were of high difficulty, and individuals may have had little knowledge of the true values. According to the theory of collective accuracy, a collective in which only a small percentage of the members have accurate knowledge could still make accurate collective judgments, because the errors of others in the collective without knowledge could cancel. However, this cancelling of errors did not occur in Experiment 1. With no knowledge of the true value, the true value did not serve as an anchor around which judgments would be randomly distributed. Experiment 1 indicated that it was more important to select a more knowledgeable subset of the collective than to use the entire collective. When the entire collective is knowledgeable, selecting the entire collective may lead to accuracy, but if the entire collective is not knowledgeable, then selecting a knowledgeable subset may lead to greater accuracy. This leads to hypothesis 5:

Hypothesis 5: Knowledge will moderate the relationship between the size and accuracy of a collective. For collectives with high knowledge, the entire collective will be more accurate than the highest ranked individual, but for collectives with low knowledge, the highest ranked individual will be more accurate than the entire collective.

Effects of variance

One of the key concepts in the theory of collective accuracy is that a collective that produces judgments with high variance is more likely to be accurate than a collective that

produces judgments with low variance because this variance will lead to a cancelling of error (Krogh & Vedelsby, 1995; Lorenz et al., 2011; Page, 2007; Surowiecki, 2005). Experiment 1 indicated that this was not the case, but the evidence was not very strong. Therefore, Experiment 2 attempts to replicate results from Experiment 1:

Hypothesis 6: The accuracy of a collective will be negatively correlated with the variance of the individual judgments.

Selecting accurate judges

If the results of Experiment 1 are valid, and it is important to select a subset from a collective to make judgments, then the method of identifying this subset is important. The process of selecting individuals for optimal performance is a major topic in industrial and organizational psychology. There are several methods available for selecting employees (Chamorro-Premuzic & Furnham, 2010). A large meta-analysis (Schmidt & Hunter, 1998) indicated that general mental ability is the best predictor of job performance. Integrity tests added the most incremental validity to the mental ability measure, and work sample tests added the second most incremental validity. Work sample tests are direct tests of tasks performed on the job. When work sample tests and mental ability are used in a single regression equation predicting performance, work sample tests have a higher standardized regression weight than mental ability. This result indicates that knowledge may be even more important than mental ability. After accounting for mental ability, the next best predictors of performance are conscientiousness and job knowledge tests. These results further support the hypothesis that knowledge is one of the most important predictors of performance in several areas.

Performance in judgment is not necessarily the same as job performance, but it is possible that selecting individuals to perform well on a judgment task is similar to selecting an individual to perform well on the job. For example, Weaver and Stewart (2011) found that three measures of intelligence were all significantly correlated with performance on a wide variety of judgment tasks. It is possible that other factors identified by Schmidt and Hunter (1998) could be used to select accurate judges, just as they were used to select high performing employees. Two predictors of job performance, integrity tests and conscientiousness, are predictors of counter-productive workplace behaviors, and therefore may not predict performance on judgment tests. This leaves mental ability tests, tests of knowledge, and performance on similar judgment tasks (the equivalent to work sample tests) as potential predictors of judgment accuracy. Experiment 1 suggested that knowledge may be the best predictor of collective accuracy. It is difficult to make a judgment without the requisite knowledge. Following knowledge, the second best predictor suggested by previous studies (Onkal, Yates, Simga-Mugan, & Oztin, 2003) is how well individuals performed on similar judgments. This is a direct, parsimonious way of detecting skill. If someone has performed well in the past, it is reasonable to suggest that they may also perform well in the future. Experiment 1 indicated that confidence does appear to be a valid indicator of judgment accuracy, but may not be as predictive as knowledge or prior performance. These facts lead to:

Hypothesis 7: The best predictors of judgment accuracy will be, in the following order: Knowledge of the domain being forecasted, skill on similar forecasting tasks, and confidence.

Using information cues to make judgments

Rather than diversity or independence, knowledge may be the key to accurate collective judgment. In Experiment 2 knowledge will be manipulated by varying the number of information cues provided to participants. Judgment is sometimes conceptualized as a process of using information cues to predict a specific value or category membership (Brunswik, 1952; Brunswik, 1956; Hammond, 1955). In artificial intelligence, statistics, and psychology, information cues are often used to predict an outcome. The idea of using variables to predict outcomes is behind the general linear model in statistics and the lens model in psychology (Hammond, 1955). The lens model has been applied in many domains, including industrial and organizational psychology (Dalal, Diab, Balzer, & Doherty, 2010). The lens model has been used to study how employees are selected (Roose & Doherty, 1976) and how nurses make workplace judgments (Holzworth & Wills, 1999).

Forecasting is a desirable task for studying use of information cues because past values can be combined in several different ways to create varied information cues. For example, Sanders (1997) provided individuals with 48 prior values of a simulated time series. In one condition participants were also provided with additional information as to the level of noise in the data and the trend and seasonality of the time series. Results showed that participants who received the additional information were more accurate than those who received only the time series.

A recent meta-analysis (Karelaia & Hogarth, 2008) indicated that across several domains, individuals could predict outcomes using cues with an average of 30% accuracy, based on R^2. Important for the current project, the meta-analysis indicated that novices ($R^2 = .31$) were more

accurate than experts (R^2 = .24), but not significantly more accurate. Accuracy decreased as a function of cues used: the R^2 for 2-cues was .40, for three cues it was .30 and for more than three cues the value was .26. This is a common theme in forecasting (Goldstein & Gigerenzer, 2009; Makridakis, Hogarth, & Gaba, 2009; Taleb, 2007), that overconfidence and a belief in the efficacy of complex over simple models leads to inaccuracy. Researchers (Gigerenzer, Todd, & ABC Research Group, 1999) have shown that simple heuristics can lead to great accuracy. However, giving participants as many cues as possible may enhance their knowledge, which was shown to be important in Experiment 1:

Hypothesis 8: Participants given the following cues will rank least to most accurate: 0-cues, 2-cues, 4-cues.

In a previous study (Ryan & Holzworth, 2012) providing participants with knowledge led to increased accuracy. Individuals were accurate at weighing information cues if they were provided only relevant cues, but were not accurate at determining which cues were relevant. One way to improve accuracy would be to give participants only relevant cues. However, in most real-world judgment and forecasting tasks, individuals must decide which cues are relevant on their own. Professional forecasters tend to study past data in order to determine which cues are relevant. Therefore, an ideal way of discovering the relevance of cues is to provide a short time series consisting of cues and outcomes. In order for individuals to carefully consider each cue, cues should be presented separately for each time period. This presentation of cues represents a short training sequence in which participants learn which cues are relevant. One interesting question is how many training trials are required for individuals to be able to identify relevant cues and properly weigh cues. In general, more learning should be expected the more trials are presented:

Hypothesis 9: Individuals who are given five training trials (five sets of cues and outcomes) will be more accurate than those given two, and those given two training trials will be more accurate than those given zero.

Effects of confidence

Aside from selecting individuals and providing individuals with knowledge, another method to improve collective judgment accuracy is to weigh individuals based on self-reported confidence. However, for this technique to be valid, individuals must be able to estimate their confidence accurately. Individuals have been shown to be overconfident in rating their judgments (Lawrence et al., 2006). This tendency makes it difficult to weigh individuals based on confidence when combining judgments. One study (Koriat, Lichtenstein, & Fischhoff, 1980) indicated that having individuals think of reasons why their estimate might be wrong led to reduced overconfidence. This technique will be tested in the current study in an attempt to reduce overconfidence:

Hypothesis 10: Collective judgments weighted by confidence will be more accurate when individuals are given instructions on how to accurately estimate confidence than when they are not given these instructions.

Method

Participants. Participants were 635 (69% women) undergraduates from a large university in the northeastern U.S. The most common major course of study was psychology with 6% of all participants majoring in psychology. They participated in exchange for partial course credit. The most accurate participant from each condition was entered into a random drawing to win one of two $150.00 gift certificates. Participants were informed that they were

not required to participate in the study and had the right to stop participating at any time. The median study completion time was 25 minutes. Participants made predictions during the early Fall of 2012.

Design. The experiment was a 4 (cues: 0, 2 recent, 2 long, 4) x 3 (training trials: 0, 2, 5) x 2 (confidence method: overconfidence information, no overconfidence information) completely crossed between subjects design, resulting in 24 conditions (see Table 9).

Prior Knowledge. Before making predictions, participants were asked a series of questions to assess how much knowledge they possessed concerning the prediction domains. They were asked:

- How much does ground beef cost per pound?
- How much does regular unleaded gas cost per gallon?
- What was the unemployment rate in percent in August of this year?
- How many individuals who were looking for their first job were still unemployed last August (answer in millions)?
- What is the average yearly GDP growth over the last 20 years?
- How much is the national average home price?
- How much did the stock price of General Electric increase per year over the last 20 years?
- How many touchdowns did Michael Turner score last season (2011 season)?

Target events. Participants were asked to predict future values of 17 target events. The outcomes were chosen to include familiar and unfamiliar outcomes. Participants were asked to predict one future value (and in some training conditions past values) for each of the following economic quantities (all are for the U.S.A.; each subdomain separated by commas represents a separate prediction):

- Ground beef prices
- Gas prices
- Unemployment rate
- Total number of individuals who are unemployed after recently entering the workforce
- GDP percentage change
- Home asking prices in the northeast, midwest, south, and east
- Stock value of 3M, Apple, GE, and Microsoft
- Average touchdowns per game scored by the National Football League running backs Steven Jackson, Michael Turner, Willis McGahee, and Frank Gore.

Cues. Most of these data are released monthly. For the monthly data, in the 4-cues condition the following cues were provided: the value one month previous, the value two months previous, the average of the entire previous year, and the overall average from the previous 20 years. For example, for ground beef prices:

September, 2012 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
?	3.45	3.45	3.32	2.32

In the 2-cues recent condition participants were provided the value one month previous and the value two months previous. For example:

September, 2012 Value	1 month before	2 months before
?	3.71	3.45

In the 2-cues long condition participants were provided with the average of the entire previous year and the overall average from the previous 20 years. For example:

September, 2012 Value	Average of entire previous year	Overall average from the past 20 years
?	3.59	1.91

In the 0-cue condition participants were not provided with cues. For example:

September, 2012 Value
?

Training trials. In the 2-trials condition participants were asked to estimate the September, 2010 and September, 2011 values before estimating the future value for September, 2012. Following each training trial participants were provided with the correct answer. For example, in the 2-trials, 4-cues condition the following would be displayed:

First page, training trial 1:

For the next group of questions you will be predicting the national average price of ground beef per pound.

Ground beef prices in 2010:

September, 2010 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
?	2.85	2.94	2.86	2.20

Your estimate for the 2010 value

Next page, feedback:

The real answer is:

September, 2010 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
2.92	2.85	2.94	2.86	2.20

Next page, training trial 2:

Ground beef prices in 2011:

September, 2011 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
?	3.23	3.27	3.10	2.25

Your estimate for the 2011 value

Next page, feedback:

The real answer is:

September, 2011 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
3.11	3.23	3.27	3.10	2.25

Next page, future prediction:

Future prediction:

Ground beef prices in 2012:

September, 2012 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
?	3.45	3.45	3.32	2.32

Your estimate for the 2012 value

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

0-100

In the 5-trials condition participants were asked to estimate the September, 2007, September, 2008, September, 2009, September, 2010, and September, 2011 values during training before estimating the future value for September, 2012. In the 0-trials condition participants were not given any prior training trials, they only estimated the value for September 2012.

Quarterly data. Data for some of the prediction domains were released quarterly rather than monthly. For the quarterly data, in the 4-cues condition, participants were provided the value from the previous quarter, two quarters previous, average of the previous year, and overall average from the previous 20 years. In the 2-cues recent condition participants were provided

the value of the previous quarter and two quarters previous. In the 2-cues long condition participants were provided with average of the previous year and overall average from the previous 20 years.

For quarterly data, in the 5-trials condition participants were asked to estimate the third quarter, 2007, third quarter, 2008, third quarter, 2009, third quarter, 2010, and third quarter, 2011 values during training before estimating the future value for third quarter, 2012. In the 2-trials condition participants were asked to estimate the third quarter, 2010 and third quarter, 2011 values.

Football predictions. Participants were asked to predict one future value (and in some training conditions past values) for each of the following National Football League quantities: Average touchdowns per game scored by the running backs Steven Jackson, Michael Turner, Willis McGahee, and Frank Gore.

In the 4-cues condition participants were provided the touchdowns from the previous year, touchdowns from two years previous, player's career average, and the average of the top 30 running backs from the previous year. In the 2-cues recent condition participants were provided with the number of touchdowns from the previous year and the number of touchdowns from two years previous. In the 2-cues long condition participants were provided with the player's career average and the average of the top 30 running backs from the previous year.

In the 5-trials condition participants were asked to estimate the average touchdowns per game for the regular season during 2007, 2008, 2009, 2010 and 2011 before estimating the future value for the 2012 season. In the 2-trials condition participants were asked to estimate the

average touchdowns for the 2010 and 2011 seasons. Following each training trial participants were provided with the correct answer.

Measures. Measures are described below and displayed at the end of Appendix B.

Demographic and individual difference measures. Participants were asked their gender, major, and verbal and math SAT scores. Sixteen other items were measured on 5-point Likert scales anchored from strongly disagree to strongly agree. Three items were taken from Tetlock (2005) regarding the hedgehog-fox dimension, faith in free markets, and optimism about the world economy. A single-item happiness measure was taken from Lykken and Tellegen (1996). Conscientiousness was measured with two items from a short measure of the Big Five Personality Domains (Gosling, Rentfrow, & Swann, 2003). The other measures were created by the author. They ask about control over events, political affiliation, and optimism about the future (see Appendix B).

Knowledge of domains. Participants were asked on a 5-point Likert scale how knowledgeable they were of the two domains queried in this study: economics and professional football. They were also asked one multiple choice question about each of these areas, "What position does Adrian Peterson play?" and "Which economist first came up with the idea of stimulating the economy through spending?"

Effort. Participants were asked on a 5-point Likert scale the extent to which they put a lot of effort into the study and the extent to which the predictions they made in the study were accurate.

Confidence. To measure confidence in their predictions participants were asked: "What is your percentage of confidence in the previous prediction? Please enter a number from 0 to 100."

Procedure. See Appendix B for a limited example of the confidence information, 4-cues, 2-trials condition. This is not a full sample, it only includes a single target event. The study was performed entirely online. Participants took the survey at the location of their choice. After registering for the experiment on a university website, participants were emailed a link and told that they had 24 hours to take the survey. This email also contained an information sheet informing participants that they have the right to withdraw from the study at any time. Participants were given a three digit code to enter when they started the survey so that they could be given partial course credit and so that they could be awarded the incentive if they made the most accurate predictions. After all incentives were awarded these codes were destroyed, resulting in anonymous data. Participants took the survey using the online survey tool *Qualtrics*©.

Participants were first told to enter the code they were emailed so that they could be given partial course credit. They were then asked what type of gift certificate they would like if they were the most accurate. Participants then answered the questions concerning background knowledge of the domains predicted.

Because the primary goal of this study is to test whether individuals can be made as accurate as possible, individuals were given brief information on how to forecast. If assigned to a condition that was not given cues, participants were told:

This study is going to ask you to make predictions about the value of certain future events (unemployment rate, number of touchdowns thrown by Tom Brady, etc.). In some cases you may learn more after making a few guesses.

If assigned to a condition given cues, participants were told:

This study is going to ask you to make predictions about the value of certain future events (unemployment rate, number of touchdowns thrown by Tom Brady, etc.). You will be given a series of variables that you can use to predict that value (see table below). These are all real data. Your job is to try to use this information, along with any personal knowledge you might have, to try to predict the next value.

In some cases all of the variables may be useful, in some cases only a few may be useful, and in some cases there won't be much of a pattern at all. If there is no pattern at all, it may be wise to simply use the long term average, because that may be the best guess. In some cases you may learn more after making a few guesses.

As a simple example, below is the U.S. population in millions:

September, 2012 value:	1 month before	2 months before	average of entire previous year	overall average from the previous 20 years
?	314.1	313.9	312.1	272.6

the guess for August 2012 might be around 314.3, because there seems to be a minor upward trend.

Next, based on previous research (Koriat et al., 1980), in the overconfidence information condition participants were told: "Individuals tend to be overconfident in their estimates. One way to avoid overconfidence is to think of reasons why your estimate may not be correct. The gift certificates will be awarded based on accuracy of predictions and accuracy of confidence ratings. Please rate your confidence carefully." In the no confidence information condition this information were omitted.

Next participants were told "Please type your answers into the small boxes on the following pages. Please just use numbers, do not use symbols like % or $. Please do not look up any additional information when making your predictions."

Participants then made predictions and completed all of the individual difference measures.

Results

Fifteen participants were eliminated after completing only the first 50% of the total questions or less. Twelve participants were eliminated for participating in the study more than once. Three participants were eliminated because the majority of their answers were extreme outliers. All variables were sorted and examined for outliers. Visual inspection was used to determine if any values that were in the far tails of the distributions. Visual inspection determined that values that were 20 standard deviations from the mean appeared to very far from the center of the distribution. Therefore these values were removed before analysis began.

Combining conditions. The 0-cues, 0-trials, overconfidence information and the 0-cues, 0-trials, no overconfidence conditions were not given any information concerning past values of the events participants predicted. These two conditions are referred to as the control conditions.

All of the 22 other conditions were given between 2 and 29 prior values of target events. These 22 conditions are referred to as the experimental conditions. In order to compare those given no knowledge to those given knowledge, for some analyses the two control conditions were combined and the 22 experimental conditions were combined and these two combined conditions were compared with one another.

Accuracy of control and experimental conditions. Table 10 displays mean judgments and errors for each target event as a function of the combined control and combined experimental conditions. As in Experiment 1, participants given no information (the control conditions) did not approach 0% error. The mean percentage error from all target events combined was 573% for the control condition. The 95% confidence interval ranged from 129% to 1017%, indicating that mean error was significantly greater than zero. In the experimental conditions, the mean percentage error was much lower, at 41%. At the mean, the experimental conditions were 14 times more accurate than the control conditions. The strongest form of the collective accuracy hypothesis would suggest that the accuracy of a random collective of individuals would be equal to the accuracy of a collective of individuals who are provided knowledge to aid judgment. At the collective level, for 11 out of 16 events the control group was significantly less accurate than the experimental groups at the $p < .05$ level. This is consistent with the idea that knowledge, rather than independence or diversity, is an important predictor of collective accuracy.

Selecting accurate subsets. Hypothesis 5 stated that knowledge will moderate the relationship between the size and accuracy of a collective. For collectives with high knowledge, the entire collective will be more accurate than the highest ranked individual, but for collectives with low knowledge, the highest ranked individual will be more accurate than the entire

collective. For this analysis the two control conditions were combined and the 22 experimental conditions were combined and these two combined conditions were compared with one another. As in Experiment 1, individuals were ranked based on accuracy on similar predictions (see Experiment 1 results section). For the current experiment, similar predictions were only used if the similar prediction significantly correlated with the prediction in both the combined control and combined experimental conditions. Replicating the procedure in Experiment 1, accuracy was computed for the highest ranked individual, highest three ranked individuals, and the entire collective.

Table 11 displays the similar prediction used to rank judgment accuracy, the correlation between the target event accuracy and the similar prediction accuracy, and the collective errors for the highest, highest three, and all participants in the control and experimental conditions. Replicating the analysis from Experiment 1, the accuracy levels were compared for the highest ranked individual, the highest three ranked, and all participants in that condition. In the control condition, there is an advantage to selecting more accurate individuals to be included in the judgment. There was only 1 of 15 cases in which the entire collective was more accurate than the highest or highest three collective mean accuracies. This result may have occurred because the correlations between judgment and similar judgment accuracy were high, indicating that there was a large difference between the most and least accurate individuals. In contrast to the theory of collective accuracy, the entire collective was rarely more accurate than using only one or three participants.

As hypothesized, the results were different in the experimental conditions. In these conditions accuracy was higher than in the control conditions and the correlations between target event accuracy and similar prediction accuracy was lower than in the control condition, so it

should be expected that selecting the best performers would not have as much of an effect on accuracy as it did in the control condition. In the control condition only one of the judgments based on the entire sample size was the most accurate compared to the highest and highest three ranked judgments, but in the experimental conditions then entire collective was the most accurate for 7 of the 15 judgments.

In order to directly test Hypothesis 5, the proportion of times the entire collective led to the most accurate judgments was compared in the experimental and control conditions. Hypothesis 5 claims that this proportion would be different in the two conditions, because in the control condition selecting the entire collective would not lead to accuracy, but in the experimental condition selecting the entire collective would lead to accuracy. Fisher's exact test indicated that these proportions did differ in the control and experimental conditions, $p = .04$. The size of this difference was large. In the control condition the entire collective was the most accurate technique 7% of the time, but in the experimental condition the collective was the most accurate technique 47% of the time.

Variances and accuracy. Hypothesis 6 stated that the accuracy of a collective will be negatively correlated with the variance of the judgments. Table 12 displays the individual accuracy and standard deviations in the control conditions and all of the combined experimental conditions. For every target event, accuracy is higher for the experimental conditions than the control conditions at the $p < .05$ level, but the variance is significantly higher in the control conditions than the experimental conditions at the $p < .001$ level. This result indicates that higher variance was associated with less accuracy.

To further explore the relationship between accuracy and variance, correlations between collective error and variance in each different condition were computed for each target event separately (see Table 13). Only the experimental conditions were included, because the control conditions would have been an overly influential data point. Mean errors and standard deviations were computed for the 22 experimental conditions, and then correlations between these 22 error and standard deviation values were computed. A total of 7 of the 17 correlations between error and standard deviation were significant, and all of these were positive. This is further evidence that low variance tends to be positively associated with accuracy.

With a positive relationship between variance and accuracy, it is possible that an accurate group may be selected by searching for a collection of individuals with the lowest variance. This could be accomplished by using the mode value of the judgments from a collective of individuals, because by definition the mode is a set of values that has zero variance. Table 14 displays the collective judgment error based on the mean, median, and mode judgment for each of the 17 target events. In the control group the mode was the more accurate than the mean and median for 71% (12 of 17) of the target events. This value is significantly higher than the 33% expected by chance, as tested with the binomial sign test, $p = .002$. In the experimental group the mode was more accurate in 18% of cases, which is not significantly different from the 33% that would be expected by chance. These data indicate that in some cases an accurate estimate can be obtained simply by selecting a subset of a collective that has low variance.

Predictors of accuracy. Hypothesis 7 states that the best predictors of judgment accuracy will be, in the following order: Knowledge of the domain being forecasted, skill on similar forecasting tasks, and confidence. Skill on similar forecasts was defined as the accuracy of the previous year (2011) prediction. Table 15 displays standardized regression coefficients for

knowledge of the domain being forecast, skill on similar forecasting tasks, and confidence in forecasts, each with forecast accuracy as the dependent variable. Skill on other tasks is significant in 14 out of 17 instances, knowledge is significant 2 of 17 instances, and confidence is significant in only 1 out of 17 instances. Hypothesis 7 was not supported, the best predictor of accuracy is skill on other forecasting tasks, with knowledge and confidence as weak predictors of accuracy.

Individual difference measures. Several individual difference measures were included as exploratory measures. Most of these had low and non-significant correlations with accuracy. The most accurate predictor of accuracy was the question "I am good at coming up with explanations for why things have occurred," which was significantly ($p < .05$) correlated with two prediction domain accuracy measures in the experimental condition and three prediction domain accuracy measures in the control condition.

Use of information cues. In order to test Hypotheses 8 and 9 regarding group differences, the 17 individual forecasting accuracy scores were standardized and then averaged. First each individual judgment was subtracted from the true score and the absolute value was taken to create an individual accuracy score for each of the 17 target events. These 17 accuracy scores were then standardized separately so that each score would be scaled in a similar manner before they were averaged. Finally the 17 standardized accuracy scores were averaged using the arithmetic mean to create one overall accuracy score per participant. The means of these overall accuracy scores as a function of condition are displayed in Table 16. A 3 (trials: 0, 2, 5) x 4 (cues: zero, 2 recent, 2 long, 4) x 2 (overconfidence information, no overconfidence information) ANOVA was performed on the standardized average accuracy scores. The results of the ANOVA are displayed in Table 17. Because variances in the groups were heterogeneous, as

indicated by Levene's test, $F(23, 573) = 37.43$, $p < .001$, the inverse of the variance of each condition was used as a weight in weighted least squares (Moder, 2010).

Hypothesis 8 states that the following groups will rank least to most accurate: 0-cues, 2-cues, 4-cues. The main effect for cues was significant, $F(3, 573) = 41.40$, $p < .001$. Planned comparisons revealed that the 0-cues condition was significantly less accurate than the 2-cues recent condition, $t(573) = 10.79$, $p < .001$, the 2-cues long condition, $t(573) = 10.77$, $p < .001$, and the 4-cues condition, $t(573) = 11.08$, $p < .001$. However, the 4-cues condition was not significantly more accurate than either the 2-cues recent condition, $t(573) = 1.50$, $p = .13$, or the 2-cues long condition, $t(573) = 1.33$, $p = .16$. In summary, the 0-cues condition was less accurate than the 2-cues recent, 2-cues long, and 4-cues conditions, but the 2-cues recent, 2-cues long, and 4-cues conditions did not significantly differ from one another.

Hypothesis 9 states that individuals who are given five training trials will be more accurate than those given two, and those given two training trials will be more accurate than those given zero. The main effect for trials was significant, $F(2, 573) = 69.05$, $p < .001$. In partial support of Hypothesis 9, the 2-trials condition was more accurate than the 0-trials condition, $t(573) = 6.28$, $p < .001$. Contrary to Hypothesis 9, the 2-trials condition was significantly more accurate than the 5-trials conditions, $t(573) = 2.08$, $p = .04$. In summary, the order of accuracy from highest to lowest was: 2-trials, 5-trials, 0-trials.

The ANOVA also indicated an interaction between trials and cues, $F(6, 573) = 18.09$, $p < .001$. This interaction is driven by the fact that the 0-cues 0-trials condition is so much less accurate than any other condition (see Table 16). With such an inaccurate condition, all mean

differences involving this condition will be much greater than any other mean difference, resulting in a significant interaction.

The ANOVA also indicated a significant main effect for the confidence, $F(1, 573) = 4.39$, $p = .04$. The confidence information condition ($M = .05$) was more accurate than the no confidence information ($M = -.02$). This effect was not hypothesized. Hypothesis 10 states that collective judgments weighted by confidence will be more accurate when individuals are given instructions on how to accurately estimate confidence than when they are not given these instructions. This hypothesis only refers to weighted confidence, there was no hypothesis concerning a difference between the confidence information and no confidence information conditions with respect to unweighted accuracy. To test Hypothesis 10, participants' accuracy scores were weighed by confidence by first summing each participant's confidence score and then dividing each score by the sum. This standardizes the score so that individuals who provide higher mean confidence do not have higher weighted accuracy scores. These confidence scores were then multiplied by the accuracy score and summed for each participant, creating an accuracy score weighted by confidence. A t-test indicated that those in the overconfidence information condition ($M = .02$) did not display more weighted accuracy than those in the no overconfidence information condition ($M = .00$), $t(518) = .62$, $p = .54$. In summary, participants in the confidence information condition were significantly more accurate than those in the no confidence information condition, but participants in the confidence information condition were not more accurate than those in the no confidence information condition with respect to confidence-weighted accuracy.

Discussion

The theory of collective accuracy claims that a collective will make judgments that approach 100% accuracy even when most of the individuals in the collective have little knowledge, because those with little knowledge will guess randomly and their errors will cancel. Replicating Experiment 1, the results of Experiment 2 indicated that a sample of individuals with little knowledge, the control conditions, did not display accuracy approaching 100%. The mean control conditions collective judgments had an error of 573%. In contrast, individuals provided with knowledge had an error rate of 41%, which is 14 times more accurate than those not provided with knowledge. This is robust evidence against the strongest form of the collective accuracy hypothesis, which states that a random sample of individuals, selected whether the individuals are knowledgeable or not, will make judgments approaching 100% accuracy. This difference between experimental and control conditions is important because it indicates that knowledge, rather than diversity or independence, is the best predictor of collective accuracy.

Contrary to the theory of collective accuracy, Experiment 1 indicated that if individuals can be ranked based on accuracy, the most accurate judgment emerges from the highest or highest three ranked individuals, rather than the entire collective. Experiment 2 replicated this result for the control conditions, in which the entire collective made the most accurate judgments in on only 1 of 15 instances. This may seem intuitive because using the most accurate individuals should result in the most accurate collective judgment. However, Experiment 2 indicated that although there were instances in which participants could be ranked based on accuracy, taking the entire collective still led to the most accuracy. In the experimental conditions, in 7 of 15 tests, taking the accuracy of the entire collective was more accurate than taking the highest or highest three ranked individuals. This was the first result from either

Experiment 1 or Experiment 2 that supported the theory of collective accuracy. This result is important because it contradicts the intuitive belief that once individuals are ranked on accuracy, we need only take the highest ranked individuals.

Variance of judgments is an important component in collective judgment theory, with several researchers (Krogh & Vedelsby, 1995; Lorenz et al., 2011; Page, 2011; Surowiecki, 2005) suggesting that high variance is required for a collective to be accurate, or that higher variance is associated with higher accuracy. In the present study, Experiment 1 indicated that lower variance of collective judgments were associated with greater accuracy. Experiment 2 replicated this result, with several results indicating that lower variance is associated with higher collective accuracy. One interesting result with clear practical implications indicated that an accurate subset of the collective was selected simply by identifying a subset with low variance, the mode judgment.

Contrary to Hypothesis 7, skill on similar forecasts was the best predictor of judgment accuracy. This evidence supports the idea that one of the best predictors of performance is performance on similar tasks. Knowledge and confidence did not predict judgment accuracy beyond skill on similar forecasts. This may have occurred because skill on similar forecasts was such a strong predictor that there was little variance left over to predict.

Most of the individual difference measures did not predict accuracy. The best predictor was the question "I am good at coming up with explanations for why things have occurred." This result implies that individuals are valid judges of their own forecasting accuracy. It is difficult to know whether individuals went into the experiment knowing that they were accurate forecasters, or observed that they were accurate after completing the predictions in the study.

This measure was actually hypothesized to be negatively correlated with accuracy, because individuals often see patterns where they do not exist and are overconfident in the validity of these patterns (Kahneman, 2011).

The only significant effect of providing cues was that providing two or four cues led to more accuracy than providing no cues. Providing individuals with more information did lead to greater accuracy, but the exact number of cues did not appear to significantly affect accuracy. The number of training trials provided affected accuracy, but not as hypothesized. As claimed in Hypothesis 9, the least accurate participants were those that performed no training trials, but contrary to Hypothesis 9, participants who completed two training trials were more accurate than those that completed five training trials. This is consistent with the idea suggested by several forecasting and judgment researchers (Makridakis, Hogarth, & Gaba, 2010b; Taleb, 2007; Yates, 1991) that individuals tend to see a pattern that does not exist and extrapolate that pattern into the future when it would be more accurate to make a forecast that is similar to the last period observed. Giving individuals more data may have allowed them to see a pattern that did not exist, which led to inaccuracy. This phenomenon of seeing a pattern that does not exist may also explain the results regarding confidence. Those who were warned that individuals tend to be overconfident were more accurate than those who did not receive this information. This information may have led participants to make more "modest" predictions, i.e. predictions that did not stray very far from the last available time period.

General Discussion

The primary purpose of this project was to test hypotheses derived from the theory of collective accuracy. Not only were most of these hypotheses not supported, but alternative

hypotheses suggesting opposite results were supported. It is important to consider why there was such a disconnect between previous results and the results from the current study.

The accuracy of large collectives

One of the strongest forms of the theory of collective accuracy suggests that a random sample of individuals, not selected based on expertise, will display accuracy approaching zero error (Lorenz et al., 2011; Surowiecki, 2005). The two experiments described in this paper provided strong evidence that this was not the case. Collectives that were not provided with knowledge had very large errors, often several orders of magnitude too high or low. There are several reasons why the current results contradict previous research.

The primary reason why these results did not replicate previous results indicating near perfect accuracy is because previous studies (Galton, 1907; Stroop, 1932) tended to use a specific type of judgment task. Studies such as these used basic perceptual tasks, such as guessing a weight, size, or temperature. These tasks represent only one type of judgment, basic perceptual judgments of quantities. A representative sample of judgment tasks is required if results are to be generalized to most judgment tasks (Brunswik, 1956). Taking a very specific subset of all judgment tasks does not tell us what will occur on other judgment tasks. This is similar to the argument Gigerenzer (1996) makes against some of the bias and heuristics research. The bias and heuristics research shows specific examples in which individuals make errors, but this does not tell us very much about how often individuals will make errors in most real-world tasks.

Perceptual judgments differ from other types of judgments in several ways. The first is that individuals often have a great deal of experience making perceptual judgments. For

example, individuals are likely to have a great deal of experience in guessing the weight of objects. This was demonstrated in the study in which a collective of individuals were 99.9% accurate in judging the weight of an ox (Galton, 1907). In the current study estimates of gas prices were consistently the most accurate judgments, because participants had a great deal of knowledge concerning gas prices relative to the other judgment tasks they performed.

Kahneman (2011) states that individuals are very accurate when judging averages, such as the average length of a line, or the number of objects in an array. This ease comes from the fact that such judgments are performed by "System 1," the intuitive, automatic system. Kahneman (2011) states that these judgments are made quickly and easily because they are done through prototype matching. The intuitive system represents categories as a prototype or a set of typical exemplars. In the example of guessing the weight of an ox (Galton, 1907), a prototypical animal is automatically activated in memory, and this prototype is used to guess the weight. This process is both automatic and accurate, which explains why individuals were accurate at judging the weight of an ox, and accurate at other perceptual judgments.

This prototype theory (Kahneman, 2011) also explains why collectives may be accurate for perceptual judgments but not for other types of judgments. Perceptual judgments easily activate a prototype because individuals commonly encounter objects such as large animals. For the judgments made in the current study, such as GDP, no relevant prototype is activated. It is difficult to imagine that individuals without knowledge of GDP could store a "prototypical GDP." A hypothesis for future research is that collective judgments will only be accurate when a relevant, accurate prototype is activated.

Another way in which perceptual judgments are unique is that almost all of the information that is required to make an accurate judgment is provided. When forecasting future values of a quantity such as the unemployment rate, individuals do not have enough information to make an accurate forecast (Makridakis et al., 2010b). Individuals would have to know how many people are unemployed and divide that by the working population in the U.S. Before this information is released, no individual outside of the U.S. Department of Labor has this information. Even professional economists make very poor judgments concerning future values of economic indicators (Taleb, 2012). In contrast, merely looking at an ox will provide almost all of the information necessary to make an accurate judgment. The sensory cues, combined with past knowledge of how much animals of similar size weigh, is enough information to make an accurate judgment. In perceptual tasks almost all the information that individuals need to make the judgment are available to the individuals, but in many other tasks the information is not available. Using a large a collective with sufficient information leads to high reliability in the use of this information, which leads to great accuracy. In the current study, the judgment of interest was forecasting. Individuals in the control groups in the current study may not have had enough information to make accurate judgments. Without this information, the errors of the collective judgments were high.

The nature of perceptual judgments also explains the fact that in previous studies (Galton, 1907; Stroop, 1932) expertise was not required for accurate collective judgments. Given that individuals will have both experience and information when making perceptual judgments, it is difficult to find experts who will be better than a novice at making perceptual judgments. Many of the previous studies of collective judgment involved perceptual tasks for which individuals all had basically the same amount of knowledge. The knowledge simply came from individuals'

senses, and in general individuals had experience making these perceptual judgments, so everyone had approximately equal knowledge. This lack of information asymmetry may explain why a subset of experts did not have to be selected for accurate judgment. There was no asymmetry in knowledge, so a more knowledgeable group could not be selected. This was in stark contrast to many of the tasks from the current set of experiments. In the current set of experiments there was a large difference in knowledge concerning the judgment tasks. With this large difference in knowledge, it was important to select a subset of individuals with high knowledge in order to make an accurate judgment. This was shown in Experiment 2, in which individuals were all given similar knowledge. When individuals were all given the same knowledge cues, in many cases an accurate subgroup could not be selected, and the most accurate technique was to select the entire collective. This result indicated that when individuals have similar knowledge, an entire collective should be used, rather than a small knowledgeable subset.

Even if a more knowledgeable subset from the collective can be selected, it could still be the case that the collective could be as accurate as the more knowledgeable subset. The less knowledgeable individuals may guess randomly, and these random guesses would cancel. Although it is an a priori fact that if random errors cancel, larger groups will approach the true mean of the population, it is an empirical question of when random error will cancel. In almost all of the tasks in the current series of studies, random error did not cancel. The errors were systematically biased, leading to inaccuracy, no matter how large the collective. In the current set of tasks individuals without knowledge did not simply guess randomly around the true value, allowing their errors to cancel. Errors were systematically biased by either being too high or too low. Although mathematical theory tells us it is possible for errors to cancel, it is an empirical

question of how often they actually will cancel in practice. In the current study the errors did not cancel.

Research on information aggregation often considers information to be a quantity that can be summed, leading to more accurate knowledge (Hayek, 1945). Adding more individuals to a collective is meant to increase accuracy because knowledge is increased. Hayek (1945) tends to think of information as a quantity that can only add accuracy to a judgment. However, what the current experiments indicate is that adding information to a collective can lead to inaccuracy if the information is inaccurate. The addition of information to a judgment is often thought of as being unequivocally beneficial. Both independent judgments and diverse judgments are thought to increase accuracy, because both processes add unique information and perspectives to a judgment (Armstrong, 2001a; Clemen & Winkler, 1985; Lorenz et al., 2011; Page, 2007). However, it appears that researchers fail to consider that even unique information that is added to a judgment can be incorrect information, and adding unique, diverse information does not guarantee that the new information is accurate. Cognitive diversity is often praised because it adds new perspectives to a judgment (Page, 2007), but it is possible that these "new perspectives" are inaccurate perspectives.

Independent judgments

The most often cited predictor of collective accuracy is independence (Armstrong, 2001a; Lorenz et al., 2011; Page, 2007; Surowiecki, 2005). Collectives in which individuals make judgments that are independent of others in the collective are hypothesized to be more accurate than collectives in which judgments are dependent. The current study not only showed that this was not the case, but indicated that the opposite is the case. In the current study dependent

judgments were shown to be more accurate than independent judgments. Some researchers (Sunstein, 2005; Surowiecki, 2005) consider "independence" to indicate a lack of live interaction. Live interaction can lead to inaccuracy through group polarization (Stoner, 1968) and groupthink (Janis, 1982). The current experiments are not relevant to these forms of dependence because there was no live interaction. Many other researchers (Lorenz et al., 2011; Page, 2007; Surowiecki, 2005) consider the statistical form of dependence to lead to inaccuracy. However, these researchers cite a result (Clemen & Winkler, 1985) that assumes that the individuals being added to the collective have an average error of zero. When this assumption is true, then adding individuals to the collective when their judgments are independent of other judgments will increase the accuracy of the collective. However, when the average error is not zero, then it is not clear what the effect of independence will be. In the current study adding independent individuals to the collective reduced accuracy because these individuals had an average error greater than zero, and the error of the entire collective increased. When dependent judgments were instead added to the collective, the increase in error was smaller because the average error of dependent judgments was lower than the average error of independent judgments.

The mathematical reason why independence may lead to accuracy is that independence creates more information (Shannon & Weaver, 1949). For example, if 10 individuals gave independent ratings of the job performance of the President of the United States, we could calculate a mean and state that the sample size is 10. However, if ratings were provided in a live setting, and nine individuals simply repeated whatever the first person stated, the sample size would really only be one. Only one independent piece of information would be provided, so the knowledge of the entire crowd is not being used. This is why it is not appropriate to analyze

dependent data as if they were independent, because one is essentially inflating sample size, and therefore inflating Type I error (Raudenbush & Bryk, 2002). Multilevel modeling was developed to address this issue of Type I error inflation. This is the reason why some researchers (Kahneman, 2011; Lorenz et al., 2011; Makridakis et al., 2010a; Page, 2007; Surowiecki, 2005) state that independence is necessary for collective accuracy, because without independence, we do not even really have a collective at all. If everyone were to simply copy the estimate of one individual, we have the estimate of an individual, not a collective. However, if the collective is not accurate, we would not want to rely on the estimate of a collective.

Without independence, the Law of Large Numbers is not necessarily valid, and we will not necessarily approach the expected value of a population parameter. What researchers (Lorenz et al., 2011; Page, 2007) who make this argument fail to recognize is that approaching the expected value of a population does not always lead to accuracy. In the current study the participants believed that the GDP growth of the U.S. was approximately 33%, when it was actually close to 3%. In this case we do not want to rely on the Law of Large Numbers because we would be guaranteed to converge on the mean of the population, but the mean of the population is not accurate. Imagine asking 10 individuals to estimate future GDP growth in a live setting. If the first person stated "I just looked up the expected GDP growth yesterday, and the consensus estimate is 2%, so I will say 2%." The next person may then think to themselves "I have never even heard of GDP growth, so rather than taking a wild guess, I will use the same estimate as the person prior, who appears to be more knowledgeable about this subject than I am." If the other 8 individuals simply followed the first knowledgeable participant, the estimate would be more far more accurate than if they simply went with their initial guess, which

according to the current study would be approximately 33%. Dependence, not independence, will lead to accuracy in cases such as these.

Part of the justification for the independence assumption is the following argument: The Law of Large numbers requires independence, collective accuracy requires the law of large numbers, therefore collective accuracy requires independence. The error is in the first premise. Even if the Law of Large Numbers were required for collective accuracy, independence is not required for the Law of Large Numbers to be valid. The Law of Large numbers assumes independence, but that does not imply that it is invalid without independence. It only indicates that if independence is assumed, then it is guaranteed to be valid. It has been proven that under certain conditions the Law of Large Numbers will hold under dependence (Birkel, 1988). Assuming that independence is a necessary condition for the law of large numbers is an example of the logical fallacy "denying the antecedent" (Pirie, 2006). An example of this fallacy is: Assuming I am a dolphin, I am a mammal; I am not a dolphin; therefore I am not a mammal. The Law of Large Numbers can be proven true under the assumption of independence, but that does not imply it is false under dependence.

Based on the differences between dependent and independent judgments from Experiment 1, the idea that independence is required for collective accuracy appears to be false. The current experiments have shown that knowledge is the most important factor in collective judgment. If knowledge is the most important predictor of collective accuracy, then it is not surprising that dependence led to more accuracy. Dependence indicates that knowledge was shared, and with knowledge being shared, more individuals will be knowledgeable. Even though it has been shown that common rather than unique knowledge tends to be shared (Stasser & Titus, 2003), if some members of the collective begin with no knowledge, then any type of

knowledge will help them make a more accurate judgment. With a more knowledgeable collective of individuals, we would expect the collective judgment to be more accurate. Therefore dependence can lead to more accuracy through the increase of knowledge in the collective. However, it is important to note that in the current study individuals did not interact in a live setting. A live setting may have been more likely to produce negative effects such as groupthink and group polarization. There may be an ideal point between complete independence and live interaction. This ideal point may involve sharing as much information as possible without having to interact in a live meeting. Future research is needed to determine where this ideal point between complete independence and live interaction lies.

Part of the confusion about independence may rely on the causal directions that are being considered. Based on the current study, if one were told that a collective made a judgment independently of others in the collective, and another collective made a judgment after sharing information with the collective, one would estimate that the information sharing collective would be more accurate than the independent collective. Dependence causes accuracy. However, if one were told that a collective all came to the same conclusion independently, and another collective came to the same conclusion after sharing information, one would trust the independent collective more than the information sharing collective. This result occurs because independent agreement is strong evidence that a judgment is accurate, because it is unlikely that all individuals would come to the same judgment independently simply by chance. This is why coding free responses in surveys is done by independent judges, and the level of agreement of the judges is reported. Independent agreement is evidence of judgment accuracy, but independence does not cause judgment accuracy.

The idea that independence is necessary for collective accuracy has been so ingrained in the scientific community that several prominent researchers state it as a simple fact. Kahneman (2011) writes "However, the magic of error reduction works well only when the observations are independent and their errors uncorrelated" (p. 84). Makridakis and coauthors (2010) state "The importance of independence suddenly becomes clear when we change the rules of the pennies-in-a-jar-game. If we start showing players the results of all previous guesses, the average estimate will wander further and further from the actual sum of the jar" (p. 208). Experiment 1 did exactly as Makridakis suggested, "showed the results of all previous guesses," and this lead to an *increase* in collective accuracy. The current project has indicated that statements concerning independence being necessary for collective accuracy may be false.

Diversity of judgments

Diversity is another factor considered to predict collective accuracy. The mechanism through which diversity increases collective accuracy is similar to the mechanism proposed for independence. In both cases more information is available to be used in the judgment, and with more information a more accurate judgment may emerge. However, more information may also simply make it difficult to select which information is correct and which information is not correct. The current experiments found repeatedly that high diversity, defined as a high variance (Krogh & Vedelsby, 1995; Page, 2007), led to less accuracy. High variance may be associated with high uncertainty or high difficulty, which is associated with less accurate judgments.

In the current study low variance was used to identify a subset of a collective with high accuracy. Using the mode judgment from a collective was more accurate than using the mean or median from the collective. This strongly contradicts the idea that high variance is associated

with greater collective accuracy. The misunderstanding that high variance can lead to greater accuracy is further caution of oversimplifying the impact of mathematical formulas. A quick glance at the bias variance decomposition formula

$$\text{Bias}^2 = \text{MSE} - \text{Variance} \quad (2)$$

seems to assure that increasing variance will decrease bias (increase accuracy). However, greater variance also leads to greater MSE (Meir, 1995), which appears to not only cancel the effect of greater variance, but completely override it. The increase in MSE is even greater than the increase in variance, leading to an increase in bias (collective inaccuracy).

As with independence, some of the confusion about the effects of diversity may depend on exactly how diversity is defined. If one is told that the set of judgments themselves are diverse, then one would expect less accuracy than if the judgments were not diverse. However, if one were told that a diverse collective of individuals, who have diverse knowledge, are making a judgment, then the effect of this type of diversity will be less clear.

Judgment vs. problem solving

Another possible explanation for the mismatch between the current study and previous studies involves the breadth of the theory of collective accuracy. Many works take a very broad approach to collective accuracy, arguing simultaneously that collectives both make more accurate judgments and are better able to solve problems (Page, 2007; Sunstein, 2006; Surowiecki, 2005). The ability to be accurate and to solve problems are often considered simultaneously, and even predictive factors such as independence and diversity are considered to be predictors of both. However, the hypotheses derived from the theory of collective accuracy

may apply to problem solving more directly than they apply to judgment. A stronger case may be made for collective problem solving efficacy than collective accuracy.

Research has shown that collectives are more likely than individuals to solve problems (Shaw, 1932). Shaw found that four-person groups solved 60% of brainteaser problems they attempted, while individuals solved only 14%. Collectives are more likely to solve problems because quantity of potential solutions will result in a higher likelihood of solving a problem. If one were trying to solve an equation, and can easily input values to the equation to test the solution, then one will be more likely to solve the equation by having as many solutions as possible. A similar effect occurs with "Eureka" problems (Lorge & Solomon, 1955). Eureka problems are those that may be difficult, but whose solution is very easy to recognize once it is suggested. An example of this type of problem is a word puzzle, such as an anagram. With these tasks, the more individuals that suggest solutions, the more likely the problem is to be solved. Lorge and Solomon (1962) suggest that for Eureka tasks the probability that a group will solve a problem is equal to:

$$1 - (1 - PI)^N$$

where PI is the probability that an individual will solve the problem and N is the group size. With a group able to produce more potential solutions than individuals, it is more likely that they will produce the correct solution.

Problem solving is a case in which it is more obvious that a large collective will be efficacious when the collective is diverse and independent. As long as solutions are easy to test, one would want several diverse solutions. These solutions may be more diverse when individuals are working independently, free from the conformity that comes from interacting

with others. If a collective created several similar suggestions, the chances of solving the problem would be lower. This basic idea is used in computer science algorithms in which a difficult problem is faced. In evolution algorithms solutions are changed randomly, simply to create diversity with the hope that with several diverse solutions the correct solution will simply be found by chance.

One important difference between problem solving and collective judgment is that often in problem solving tasks the solutions can be tested in some way in order to determine whether the solutions are accurate. In collective judgment the situation is often the opposite; whether the judgment is accurate is the primary question. In problem solving generating a number of diverse solutions can increase the probability of finding the correct solution, because there are more solutions to test. In judgment, generating a number of diverse judgments simply creates more ambiguity and uncertainty about what the accurate value is.

Using a large collective may be more efficacious in problem solving than in judgment. The success associated with collective problem solving may have been over-generalized to collective judgment, resulting in an overstatement of the judgment accuracy of a large collective.

Selection of an accurate subset

Researchers (Larrick & Soll, 2006) have shown that it is not intuitive that averaging results from a large collective can increase accuracy. However, in the current study, there were several cases in which even though individuals could be ranked based on accuracy, the entire collective was still more accurate than a knowledgeable subset. In a single experiment, it was shown that when individuals had large differences in individual accuracy, then selecting a subset that was more knowledgeable led to more accuracy than using the entire subset in 14 out of 15

instances. However, when individuals did not possess large differences in knowledge, in 7 of 15 instances the collective was more accurate than a more knowledgeable subgroup. What may appear to be obvious, that selecting a more knowledge subset will lead to greater accuracy, has been shown to be wrong on theoretical grounds (Krogh & Vedelsby, 1995; Page, 2007) and was shown to be wrong on empirical grounds in the current study. This is important because it was the only result in the current project that confirmed a hypothesis derived from the theory of collective accuracy. This result confirmed the surprising prediction that adding less accurate individuals to a collective can increase the accuracy of the collective judgment.

This result also confirms the intuition that it will be important to select an accurate subset when there are large individual differences in knowledge. In the control group of Experiment 2, the correlations that were used to rank individuals ranged from .57 to .89. For these judgments, there were no cases in which the collective was more accurate than a more knowledgeable subset. In the experimental conditions, the correlations ranged from .16 to .42. For these conditions, in 7 of 15 cases the judgment using the entire collective was the most accurate. Although higher correlations made it easier to select an accurate subset of individuals, there was no simple threshold correlation that determined when the collective would be accurate and when it would not. In the experimental condition, the three highest correlations, .41, .41, and .42, were all cases in which the collective was the most accurate. The lowest correlation, .16, was a case in which a more knowledgeable subset was more accurate than the entire collective. It may not be possible to simply look at a correlation used to rank individuals on accuracy and use the strength of that correlation to determine with certainty whether the entire collective or a subset should be used. However, one should be more likely to use the entire collective when the correlation used

to rank individuals is low. Exactly how low this correlation should be is a difficult question. This question requires further theoretical and empirical investigation.

Determining truth

One of the hopes from the theory of collective accuracy is that it provides a way to verify the truth of propositions. This is especially true in cases of consensus scoring, in which correct answers are considered to be the answers most often selected by a pool of respondents (Mayer et al., 2004; Mohoric et al., 2010; Warwick et al., 2010). The results from the current experiments do not support the accuracy of consensus scoring. A large collective was shown to be inaccurate in many of the judgments from the current set of experiments. Philosophers have been considering how to determine truth for thousands of years, and surprisingly, there is general agreement on how to establish truth. The best way to determine whether a proposition is true is to use either a priori or empirical tests. A priori tests are logical or mathematical proofs. Empirical tests are tests that involve observing whether the results are true, such as the process of experimentation.

Some of the reasoning behind the theory of collective accuracy is that experts should not be trusted to establish accuracy, but rather one should rely a large collective to determine accuracy. The results from the current study, combined with other research on expertise (Meehl, 1954; Tetlock, 2005), may suggest that neither experts nor collectives should be relied upon for accuracy. This has important implications for industrial and organizational psychology. Whenever possible, a leader should use a priori or empirical knowledge to establish truth. If these methods are not available, a small group or even single individual that has a great deal of knowledge should be relied upon. The results from the current experiment indicate that a large

collective may be the last group that should be relied upon for accuracy. A large group may suggest solutions to problems, or add knowledge that a single expert may not have, but a large collective is more likely to be inaccurate than a smaller group with greater knowledge. However, if a collective can be found in which individuals all have equal knowledge, then as large a collective as possible should be used.

If a collective is being used, the current series of experiments have shown when a collective may be more likely to be accurate. Contrary to previous research on independence (Lorenz et al., 2011; Page, 2007; Surowiecki, 2005), the current experiments have shown that when information is shared in the collective, the collective may be more accurate. Also contrary to previous research (Lorenz et al., 2011; Page, 2007; Surowiecki, 2005), a collective should be trusted more when the variance of the judgments from the collective is low than when the variance is high. High variance indicates disagreement, which implies that the judgment is of high difficulty. Low variance indicates that the judgment is obvious, and more trust should be placed in the judgment.

Although estimates vary, it has been suggested that data stored on computers throughout the world is doubling at least every two years (Gantz & Reinsel, 2011). It may appear that with more data available, more scientific and scholarly progress can be made. The idea that more data are a positive factor is partly inspired by the theory of collective accuracy. With more information available, more accurate knowledge may be available. Given that the current study has cast doubt on the theory of collective accuracy, the large increase in data may be worrisome. Doubling the amount of data may create some accurate data, but it also may create some inaccurate data. Increasing data may simply make it more difficult to find the accurate data among the inaccurate data. If an Internet search produces ten results that all disagree with one

another, it is difficult to determine which of those are accurate. The Internet is often referred to as a democratizing influence (Lanier, 2013). In the past in order to broadcast to a large number of citizens one needed access to radio or television stations that cost millions of dollars. This meant that only professionals with a great deal of resources could broadcast information. With the advent of the Internet, amateurs can now broadcast information as well. This is sometimes seen as a democratizing, positive influence (Kurzweil, 1999), but it also makes it difficult to distinguish knowledgeable professionals from inaccurate novices. If a large collective tends to be accurate, regardless of their level of knowledge, then this democratizing of information distribution could be seen as a positive development. However, with the current study indicating that the collective is often inaccurate, this development may be detrimental.

Applications to Industrial and Organizational Psychology

Implications for collective judgment in organizations. Live face to face meetings have several detrimental effects. These include groupthink (Janis, 1972), group polarization (Stoner, 1968) conformity (Asch, 1951), and providing of shared rather than unique information (Stasser & Titus, 2003). With so many negative aspects, it is surprising that organizations focus so heavily on live meetings for collective judgments and decisions (Sunstein, 2006). The current study suggests that sharing information, rather than making judgments independently, leads to accuracy. These results imply that information should be shared, but ideally not shared in a live setting. The Delphi technique (Rowe & Wright, 1999) allows individuals to share information without live interaction. Markets also display this ideal mix of information sharing and independent judgment. This optimal mix may suggest why the Delphi technique and markets have been shown to be the most accurate forms of judgment (Fama, 1970; Graefe & Armstrong, 2011; Rowe & Wright, 1999). These techniques may have been utilized less than live meetings

in the past due to their difficulty in implementation, but with the availability of the Internet making these techniques easier to implement, their use may begin to increase. Another technique that allows information sharing without some of the biases of live interaction is the stepladder technique (Rogelberg et al., 1992). The stepladder technique prompts individuals to independently consider a judgment before discussing the judgment with other members of a group, and adds individuals to a group one at a time so that each individual's unique information is considered. An advantage of this technique is that it is simple and does not require technology to implement.

Implications for leadership. Vroom (2000) presents a theory of leadership that examines the conditions under which leaders should involve subordinates in decision making and judgment. The theory considers many factors relevant to organizational decision making, such as the importance of gaining support from subordinates, leader expertise, and subordinate expertise. Vroom (2000) states that when leaders are high in expertise for the particular decision and it is not important to gain commitment from subordinates, the leader should make the final decision on their own, with the possibility of gaining some input from others. The final decision is made by the leader, but the theory does not state under which conditions the leader should seek input from others. The current study has indicated that this detail, whether to gain information from others, is a very important and controversial one. The results from the current study would suggest that leaders should solicit input from others only if they have expertise that is similar to the leader.

Vroom's (2000) research suggests that important components in leader decision making move beyond mere accuracy, such as the importance of gaining commitment from subordinates, and concerns for development of employee skills by involving them in judgments and decisions.

The current project only deals with judgment accuracy. There may be cases in organizational judgment in which individuals should be included in a judgment in order to increase satisfaction and acceptance of the judgment, even if their inclusion leads to less accurate judgments. Subordinates should be included if it is more important for the judgment to be accepted than for the judgment to be accurate. These two factors, acceptability and accuracy of a judgment, must carefully be weighed in organizational judgment.

Organizational leaders may be aware of the recent literature on collective judgment and action. In fact, the idea of collective accuracy has become so popular that leaders may have heard of the research in the popular press. This popularity may instill a tendency for leaders to use a large collective to make judgments. However, if that large collective is lacking in knowledge, this approach can lead to inaccurate judgments. When leaders have a great deal of expertise, and no one can be found with similar expertise, a leader may need to make an important judgment independently of others in the organization. Another method a leader could use is to educate a large collective so that they will have the knowledge necessary to help make a judgment (Vroom, 2000).

If leaders do not have the knowledge to make a forecast themselves, they may have no choice but to look to others in the organization to help make a forecast. One way to accomplish this is to establish an internal prediction market. A prediction market allows individuals to make "bets" on future events. These bets provide individuals with the incentive to provide accurate judgments (Arrow et al., 2008). A study of an internal market at Hewlett Packard found that small internal prediction markets, with only 26 participants, predicted printer sales with significantly greater accuracy than official company forecasts (Chen & Plott, 2002).

Leaders do not only have the option of asking employees for solutions, but also going entirely outside the organization for solutions. With the advent of the Internet, the idea of letting thousands of potential solvers work on a problem is no longer a theoretical question, but a practical one. The practice of posting a problem on the Internet and asking for a solution is referred to as crowdsourcing (Howe, 2005). Innocentive.com is a site on which companies can post problems that they would like to have solved. An example problem taken from the site is:

> ***Recovering Bacillus Spores from Swabs***
>
> ***TAGS: Life Sciences, Global Health, Food/Agriculture, Environment, RTP***
>
> ***AWARD: $30,000 USD | DEADLINE: 4/05/12 | ACTIVE SOLVERS: 38 | POSTED: 1/05/12***
>
> ***The Seeker requires protocols for efficient recovery of bacterial spores (Bacillus subtilis/Bacillus atrophaeus) from pre-wetted surface sampling tools with handle. Guidance and standardized protocols are provided within the detailed challenge description.***
>
> ***This is a Reduction-to-Practice Challenge that requires a written proposal and experimental proof-of-concept data.***

When posting the problem a company specifies what it will accept as demonstration of the solution. These are sometimes replications of experimental results, or a theoretical evaluation of the proposal by the seeker (Https://Www.innocentive.com/ar/challenge/9932938.). Innocentive's founding is described by the co-founder of the company in a recent paper (Lakhani & Panetta, 2007). The idea for the company came from Alph Bingham, who was the Vice

President of Research and Development at Eli Lilly. He observed in his doctoral program that most scientific problems were amenable to multiple approaches and diverse solutions, and excelling in one area of science was not a good indicator of success in other areas. Therefore, opening the problem to individuals outside of the problem domain may lead to a solution. Lakhani and Jeppesen (2007) found that 30% of the problems posted on Innocentive were solved. They note that this is a high number given that these were problems that could not be solved by scientists working for the company that posted them.

Another situation in which a leader should rely on a large collective is when a leader is trying to solve a problem and potential solutions can be tested easily. In this case, it is a simple matter of mathematics that the more potential solutions are identified, the more likely the problem is to be solved (Lorge & Solomon, 1962). This is one instance in which a leader should seek several solutions from a diverse collective of individuals, regardless of knowledge or ability. More diverse individuals may provide more diverse solutions, covering more of the solution space and leading to more likelihood of solving the problem. This technique must be used carefully, however, because if too many incorrect solutions are proposed, identifying the correct solution may become more difficult.

Implications for personnel selection. Employee selection often involves multiple individuals, such as having several perspective employers interviewing a candidate simultaneously (panel interviews) (Dixon, Wang, Calvin, Dineen, & Tomlinson, 2002), having several perspective employers interview candidates serially (Dose, 2003), and having different interviewers provide input to the final selection decision (Campion & Palmer, 1997). All three of these areas are affected by the controversies over information sharing vs. independent aggregation of information.

A meta-analysis (McDaniel, Whetzel, Schmidt, & Maurer, 1994) indicated that panel interviews were less valid predictors of job performance than individual interviews, but the authors did not specify whether individual interviews were conducted by a single interviewer or whether selections were made based on a series of individual interviews. A review (Dixon et al., 2002) revealed that the validity of serial, panel, or single interviews was inconclusive. The same review also considered the validity of two techniques for final selection of candidates. One technique is to have individuals provide scores independently and then select the candidate with the highest average score. A second method is to have several employers meet and attempt to reach consensus. This review (Dixon et al., 2002) again found that previous research was contradictory and inconclusive with regard to the validity of these two methods.

The inconclusive nature of these results is likely to have occurred because the lack of careful consideration of the effect of information sharing vs. independence in the personnel selection process. Information sharing increases knowledge, but may create bias that independent judgments avoid. An interesting potential bias induced by information sharing is discussing candidates between interviews, before all candidates have been interviewed (Campion & Palmer, 1997). This discussion could create order effects, introduce irrelevant information, and change standards for future candidates. Information sharing also may explain the inconclusive results involving panel interviews. Panel interviews leverage a larger collective, which would suggest that they would be more valid than a single one-on-one interview due to increased reliability. However, a panel interview would suffer from the negative effects of live interaction such as groupthink (Janis, 1972) and group polarization (Stoner, 1968). A series of individual interviews may appear beneficial, but this may create too much information (Gigerenzer et al., 1999; Goldstein & Gigerenzer, 2009), and it may be difficult to combine all of

this information when the final selection is made, especially if live discussion is used. Attempting to reach consensus through discussion may also suffer from groupthink and group polarization, and also may destroy unique information through conformity and by encouraging a focus on shared rather than unique information (Stasser & Titus, 2003). For the final candidate selection, rather than using consensus, information could be exchanged, but candidates could be given independent numerical ratings.

Results from the current study suggest a compromise of sharing information but avoiding the negative biasing effects of live interaction. Candidates should be interviewed separately by several individuals. A Delphi technique (Rowe & Wright, 1999) should then be used to select candidates. Information would be posted online anonymously to a discussion board, and individuals could then provide an average rating for each candidate. Several rounds of discussion and ratings would be solicited until the ratings changes were negligible. If this full technique is not feasible, a single round of information sharing and rating could be used. The candidate with the highest mean or median rating would then be selected. Future research is necessary to test this hypothesized selection system.

Future research

Given such a discrepancy between the current and previous research, future research is needed to determine further predictors of collective accuracy. Knowledge was shown to be the key predictor of collective accuracy in the current study. When individuals possessed similar knowledge, the collective was more likely to be accurate than when there were large differences in knowledge. The key question that remains is how similar this knowledge must be in order for the collective to be more accurate than a more knowledgeable subset. There also remains a

question of how accurate a large collective will be if it contains individuals with a mix of accurate and inaccurate information.

The effect of independence is another clear avenue for future research. Again, the current results did not confirm previous results (Armstrong, 2006; Kahneman, 2011; Lorenz et al., 2011; Makridakis et al., 2010a; Sunstein, 2006). In the current study dependence resulted in the sharing of information with less knowledgeable individuals, leading to greater accuracy. Therefore it may be expected that information sharing will increase accuracy when there is a large difference in knowledge. However, when there is not a large difference in knowledge information sharing may lead to bias or group polarization. These results would imply an interaction in which information sharing increases the accuracy of judgments when there are large differences in knowledge, but decrease accuracy when there are large similarities in knowledge. This hypothesis requires further testing.

It is possible that part of the reason why the theory of collective accuracy was not supported in the current study is because the collective accuracy associated with problem solving was overgeneralized to judgment. There is strong theoretical and empirical evidence that problems are more likely to be solved when several individuals suggest solutions (Lorge & Solomon, 1962). However, given the results of the current set of experiments, it may be the case that the area of collective problem solving is similar to the area of collective judgment and that the efficacy of collective problem solving is also overstated. Collective problem solving will only be accurate when solutions can be easily tested. Even if solutions can be easily tested, it is possible that if too many solutions are tested, one may appear to solve the problem, but this may occur only be chance, as occurs in the problems of overfitting and inflation of Type I error. It could also be argued that in practical situations it is rare that solutions can be easily tested. One

of the primary ways to verify that a solution to a scientific problem is correct is through empirical tests such as experiments (Moser, 1987). If thousands of potential solutions are generated, it may not be practical to test all of them by performing thousands of experiments. We may therefore need to reduce the number of potential solutions to those that are most likely to solve the problem, and this may require specialized knowledge. This situation is similar to collective judgment, in which the current study has shown that a smaller, more knowledgeable subset should be used rather than the entire collective. Like collective judgment, the efficacy of collective problem solving may also be overstated.

Diversity is another avenue for future research. Contrary to prior research, the current set of experiments indicated that more diverse judgments were less accurate. However, there is a remaining question about the diversity of other factors in addition to the judgments themselves. It has been suggested that collectives that contain individuals with diverse knowledge or diverse perspectives may be more accurate than collectives without these diverse qualities (Armstrong, 2001a; Page, 2007). Given the strong results in the current study suggesting judgment diversity leads to inaccuracy, these additional forms of diversity may also lead to inaccuracy. This is an interesting hypothesis for future research.

Conclusion

Recent research touting the efficacy of large collectives should be viewed critically. Collectives can be very accurate, but the current experiments indicate that this accuracy may only occur when individuals in the collective all have very similar knowledge. Such a situation may be the exception in practice, indicating that high collective accuracy may be a rare phenomenon. A small set of knowledgeable experts will often be more accurate than a large

collective of individuals. It may be more important to select a knowledgeable collective than a large collective. Contrary to previous research (Lorenz et al., 2011; Page, 2007; Surowiecki, 2005), knowledge, rather than independence or diversity, is the most important factor predicting collective accuracy. With knowledge as the most important predictor of accuracy, organizations should focus on educating as many individuals as possible. Education and sharing of information at all levels is the key to organizational success.

References

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., . . . Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology. Applied, 6*(2), 130-47.

Armstrong, J. S. (2001a). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners.* (pp. 417-440) Kluwer Academic Publishers.

Armstrong, J. S. (2001b). Evaluating methods. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners.* (pp. 441-472) Kluwer Academic Publishers.

Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings. Retrieved 1/13, 2012, from http://repository.upenn.edu/marketing_papers/44

Arrow, K. J., Milgrom, P., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., . . . Zitzewitz, E. (2008). Economics: The promise of prediction markets. *Science, 320*(5878), 877-878.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow, & H. Guetzkow (Eds.), *Groups, leadership and men; research in human relations.* (pp. 177-190). Oxford England: Carnegie Press.

Birkel, T. (1988). A note on the strong law of large numbers for positively dependent random variables. *Statistics & Probability Letters, 7*(1), 17-20. doi:http://dx.doi.org/10.1016/0167-7152(88)90080-6

Brodie, M. (2012). Kaiser poll finds bipartisan support for spending on global health. Retrieved August 28, 2013, from http://kff.org/global-health-policy/press-release/kaiser-poll-finds-bipartisan-support-for-spending-on-global-health/

Brožek, J., & Tiede, K. (1952). Reliable and questionable significance in a series of statistical tests. *Psychological Bulletin, 49*(4), 339-341. doi:10.1037/h0058274

Brunswik, E. (1952). *The conceptual framework of psychology. (int. encycl. unified sci., v. 1, no. 10.)*. Oxford England: Univ. Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* Berkeley: University of California Press.

Campion, M. A., & Palmer, D. K. (1997). A review of structure in the selection interview. *Personnel Psychology, 50*(3), 655-702.

Chamorro-Premuzic, T., & Furnham, A. (2010). *The psychology of personnel selection*. Cambridge, UK; New York: Cambridge University Press.

Chen, K. Y., & Plott, C. R. (2002). *Information aggregation mechanisms: Concept, design and implementation for a sales forecasting problem.* Unpublished manuscript.

Clemen, R. T., & Winkler, R. L. (1985). Limits for the precision and value of information from dependent sources. *Operations Research, 33*(2), 427-442.

Condorcet, J. d. C. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie royale.

Dalal, D. K., Diab, D. L., Balzer, W. K., & Doherty, M. E. (2010). The lens model: An application of JDM methodologies to IOOB practice. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*(4), 424-428. doi:10.1111/j.1754-9434.2010.01263.x

Dalal, R. S., Bonaccio, S., Highhouse, S., Ilgen, D. R., Mohammed, S., & Slaughter, J. E. (2010). What if industrial-organizational psychology decided to take workplace decisions seriously? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*(4), 386-405. doi:10.1111/j.1754-9434.2010.01258.x

Davis, J. (1992). Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples 1950:1990. *Organizational Behavior and Human Decision Processes, 52*(1), 3-38.

Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology, 53*(3), 497-509.

Dixon, M., Wang, S., Calvin, J., Dineen, B., & Tomlinson, E. (2002). The panel interview: A review of empirical research and guidelines for practice. *Public Personnel Management, 31*(3), 397.

Dose, J. J. (2003). Information exchange in personnel selection decisions. *Applied Psychology: An International Review, 52*(2), 237-252. doi:10.1111/1464-0597.00133

Esser, J. (1998). Alive and well after 25 years: A review of groupthink research. *Organizational Behavior and Human Decision Processes, 73*(2\3), 116.

Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *JOURNAL OF FINANCIAL ECONOMICS -AMSTERDAM-, 49*(3), 283-306.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance, 25*(2), 383-417.

Fischer, G. (1981). When oracles fail? A comparison of four procedures for aggregating subjective probability forecasts. *Organizational Behavior and Human Performance, 28*(1), 96-110.

Galton, F. (1907). Vox populi. *Nature, 75*, 450-451.

Gantz, J., & Reinsel, D. (2011). *Extracting value from chaos.*

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to kahneman and tversky. *Psychological Review, 103*(3), 592-596. doi:10.1037/0033-295X.103.3.592

Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121*(1), 149-167.

Gigone, D., & Hastie, R. (1993). The common knowledge effect: Information sharing and group judgment. *Journal of Personality and Social Psychology, 65*(5), 959-974.

Gilson, R. J., & Kraakman, R. (2003). The mechanisms of market efficiency twenty years later: The hindsight bias. *The Journal of Corporation Law., 28*(4), 715.

Goldstein, D. G., & Gigerenzer, G. (2009). Fast and frugal forecasting. *International Journal of Forecasting, 25*(4), 760-772. doi:10.1016/j.ijforecast.2009.05.010

Goodman, B. (1972). Action selection and likelihood ratio estimation by individuals and groups. *Organizational Behavior and Human Performance, 7*(1), 121-141.

Gordon, K. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology, 7*(5), 398-400.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. J. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37*(6), 504-528. doi:10.1016/S0092-6566(03)00046-1

Graefe, A., & Armstrong, J. S. (2011). Comparing face-to-face meetings, nominal groups, delphi and prediction markets on an estimation task. *International Journal of Forecasting, 27*(1), 183-195. doi:10.1016/j.ijforecast.2010.05.004

Grinstead, C. M., & Snell, J. L. (1997). *Introduction to probability.* Providence, RI: American Mathematical Society.

Gustafson, D. H., Shukla, R. K., Delbecq, A., & Walster, G. W. (1973). A comparative study of differences in subjective likelihood estimates made by individuals, interacting groups, delphi groups, and nominal groups. *Organizational Behavior & Human Performance, 9*(2), 280-291. doi:10.1016/0030-5073(73)90052-4

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review Psychological Review, 62*(4), 255-262.

Hayek, F. A. (1945). The use of knowledge in society. *The American Economic Review, 35*(4)

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88*(2), 359-369. doi:10.1037/0033-2909.88.2.359

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin, 121*(1), 43-64.

Holzworth, R. J. (2002). *Biodata questionnaire.* Unpublished manuscript.

Holzworth, R. J., & Wills, C. E. (1999). Nurses' judgments regarding seclusion and restraint of psychiatric patients: A social judgment analysis. *Research in Nursing & Health, 22*(3), 189-201. doi:10.1002/(SICI)1098-240X(199906)22:3<189::AID-NUR2>3.0.CO;2-Q

Howe, J. (2005). The rise of crowdsourcing. *Wired, 14.06*

Https://Www.innocentive.com/ar/challenge/9932938. Retrieved 01/12/2012

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology, 50*(6), 1141-1151.

Janis, I. L. (1972). *Victims of groupthink; a psychological study of foreign-policy decisions and fiascoes*. Boston: Houghton, Mifflin.

Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes*. Boston: Houghton Mifflin.

Jensen, M. C. (1978). Some anomalous evidence regarding market efficiency. *Journal of Financial Economics, 6*(2-3), 95-101.

Kac, M. (1959). *Statistical independence in probability, analysis and number theory*. New York: Mathematical Association of America; distributed by Wiley.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404-426. doi:10.1037/0033-2909.134.3.404; 10.1037/0033-2909.134.3.404.supp (Supplemental)

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55*, 623-55.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107-118. doi:10.1037/0278-7393.6.2.107

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 7*, 231 - 238.

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence*. New York: Viking.

Lakhani, K. R., & Jeppesen, L. B. (2007). Getting unusual suspects to solve R&D puzzles. *Harvard Business Review, 85*(5), 30-32.

Lakhani, K. R., & Panetta, J. A. (2007). The principles of distributed innovation. *Innovations: Technology, Governance, Globalization, 2*(3), 97-112.

Langlois, J., & Roggman, L. (1990). Attractive faces are only average. *Psychological Science, 1*(2), 115-121.

Lanier, J. (2013). *Who owns the future?* Simon & Schuster.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *MANAGEMENT SCIENCE, 52*(1), 111-127.

Latan, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37*(6), 822-832.

Lawrence, M., Goodwin, P., O'connor, M., & Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting, 22*(3), 493-518. doi:http://www.elsevier.com/wps/find/journaldescription.cws_home/505555/description#description

Lindgren, B. W. (1993). *Statistical theory*. New York, NY: Chapman & Hall.

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

Lorge, I., Fox, D., Davitz, J., & Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychological Bulletin, 55*(6), 337-72.

Lorge, I., & Solomon, H. (1955). Two models of group behavior in the solution of eureka-type problems. *Psychometrika, 20*(2), 139-148.

Lorge, I., & Solomon, H. (1962). Group and individual behavior in free-recall verbal learning. In J. H. Criswell, H. Solomon & P. Suppes (Eds.), *Mathematical methods in small group processes.* (pp. 221-231). Stanford, CA: Stanford University Press.

Lykken, D., & Tellegen, A. (1996). Happiness is a stochastic phenomenon. *Psychological Science, 7*(3), 186-189. doi:10.1111/j.1467-9280.1996.tb00355.x

Makridakis, S., Hogarth, R. M., & Gaba, A. (2009). Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting, 25*(4), 794-812.

Makridakis, S., Hogarth, R. M., & Gaba, A. (2010a). *Dance with chance: Making luck work for you*. Richmond: Oneworld.

Makridakis, S., Hogarth, R. M., & Gaba, A. (2010b). FORECASTING - why forecasts fail. what to do instead. *MIT Sloan Management Review., 51*(2), 83.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2004). TARGET ARTICLES: "Emotional intelligence: Theory, findings, and implications". *Psychological Inquiry, 15*(3), 197-215.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*(4), 599-616. doi:10.1037/0021-9010.79.4.599

McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, N.J.: Prentice-Hall.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN US: University of Minnesota Press. doi:10.1037/11281-000

Meir, R. (1995). Bias, variance and the combination of least squares estimators. In T. Leen, G. Tesauro & D. Touretzky (Eds.), *Advances in neural information processing systems 7: Proceedings of the 1994 conference* (pp. 295-302)

Moder, K. (2010). Alternatives to F-test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling, 52*(4), 343-353.

Mohoric, T., Taksic, V., & Duran, M. (2010). In search of 'the correct answer' in an ability-based emotional intelligence (EI) test. *Studia Psychologica, 52*(3), 219-228.

Moser, P. K. (1987). *A priori knowledge*. Oxford [Oxfordshire]; New York: Oxford University Press.

Myers, D. G. (1975). Discussion-induced attitude polarization. *Human Relations Human Relations, 28*(8), 699-714.

Onkal, D., Yates, J. F., Simga-Mugan, C., & Oztin, S. (2003). Professional vs. amateur judgment accuracy: The case of foreign exchange rates. *Organizational Behavior and Human Decision Processes., 91*(2), 169.

Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton: Princeton University Press.

Page, S. E. (2011). *Diversity and complexity*. Princeton, NJ: Princeton University Press.

Paulus, P., Dugosh, K., Dzindolet, M., Coskun, H., & Putman, V. (2002). Social and cognitive influences in group brainstorming: Predicting production gains and losses. *European Review of Social Psychology, 12*(01), 299-325.

Paulus, P., & Dzindolet, M. (1993). Social influence processes in group brainstorming. *Journal of Personality and Social Psychology, 64*(4), 575-586.

Pirie, M. (2006). *How to win every argument: The use and abuse of logic*. London: Continuum.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage Publications.

Reilly, F. K., & Brown, K. C. (2009). *Investment analysis and portfolio management*. Mason, OH: Thomson/South-Western.

Rogelberg, S. G., Barnes-Farrell, J., & Lowe, C. A. (1992). The stepladder technique: An alternative group structure facilitating effective group decision making. *Journal of Applied Psychology, 77*(5), 730-737. doi:10.1037/0021-9010.77.5.730

Roose, J. E., & Doherty, M. E. (1976). Judgment theory applied to the selection of life insurance salesmen. *Organizational Behavior & Human Performance, 16*(2), 231-249.

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting, 12*(1), 73-90.

Rowe, G., & Wright, G. (1999). The delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting., 15*(4), 353.

Ryan, S., & Holzworth, R. J. (2012). *Cue use in a forecasting task.* Unpublished manuscript.

Sanders, N. R. (1997). The impact of task properties feedback on time series judgmental forecasting tasks. *Omega, 25*(2), 135.

Schmidt, F. L., & Hunter, J. E. (1998). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings.* Washington, DC: American Psychological Assoc.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

Shanteau, J., & Stewart, T. R. (1992). Why study expert decision making? some historical perspectives and comments. *Organizational Behavior and Human Decision Processes, 53*(2), 95.

Shaw, M. E. (1932). A comparison of individuals and small groups in the rational solution of complex problems. *The American Journal of Psychology, 44*(3), 491-504.

Stasser, G., & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry, 14*(3-4), 304-313. doi:10.1207/S15327965PLI1403&4_21

Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.

Stoner, J. (1968). Risky and cautious shifts in group decisions: The influence of widely held values. *Journal of Experimental Social Psychology Journal of Experimental Social Psychology, 4*(4), 442-459.

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology, 15*(5), 550-562.

Sunstein, C. R. (2005). Group judgments: Statistical means, deliberation, and information markets. *New York University Law Review., 80*(3), 962.

Sunstein, C. R. (2006). *Infotopia: How many minds produce knowledge*. Oxford; New York: Oxford University Press.

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

Taleb, N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Taleb, N. (2012). *Antifragile: Things that gain from disorder*. New York: Random House.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? how can we know?*. Princeton, N.J.: Princeton University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (New York, N.Y.), 185*(4157), 1124-31.

Vroom, V. H. (2000). Leadership and the decision-making process. *Organizational Dynamics, 28*(4), 82-94.

Warwick, J., Nettelbeck, T., & Ward, L. (2010). AEIM: A new measure and method of scoring abilities-based emotional intelligence. *Personality and Individual Differences, 48*(1), 66-71. doi:10.1016/j.paid.2009.08.018

Weaver, E., & Stewart, T. R. (2011). Dimensions of judgment: Factor analysis of individual differences. *Journal of Behavioral Decision Making,* doi:10.1002/bdm.748

Weldon, M. S., & Bellinger, K. D. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology. Learning, Memory, and Cognition., 23*(5), 1160.

Wright, D. B., & Klumpp, A. (2004). Collaborative inhibition is due to the product, not the process, of recalling in groups. *Psychonomic Bulletin & Review, 11*(6), 1080-1083.

Yammarino, F. J., & Dansereau, F. (2008). Multi-level nature of and multi-level approaches to leadership. *The Leadership Quarterly, 19*(2), 135-141. doi:10.1016/j.leaqua.2008.01.001

Yammarino, F. J., Salas, E., Serban, A., Shirreffs, K., & Shuffler, M. L. (2012). Collectivistic leadership approaches: Putting the 'we' in leadership science and practice. *Industrial & Organizational Psychology, 5*(4), 382-402. doi:10.1111/j.1754-9434.2012.01467.x

Yates, J. F. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior and Human Decision Processes, 49*(1), 60.

Table 1

Unit of Measurement, True Values, Mean Judgments, and Errors for the Control and Dependent Conditions for Sample One

			Control			Dependent		
Target Event	Unit	True Value	*M*	Error[a]	*n*	*M*	Error[a]	*n*
U.S. GDP	Percent	2.2	20.3	820%	24	22.9	940%	21
China's GDP	Percent	8.1	37.2	359%	24	32.5	301%	21
Unemployment Rate	Percent	8.1	15.9	97%	24	13.0	60%	21
Retail Sales	Percent	-0.2	13.7	6965%	24	22.2	11220%	21
Gas Price	$	3.9	4.0	4%	24	4.5	17%	21
Gold Price	Hundreds of $	16.6	7.2	56%	24	7.1	57%	21
Home Sales	Millions	5.0	0.4	92%	24	0.1	98%	21
Unemployment Claims	Millions	0.4	2.7	627%	24	0.2	48%	21
The Avengers Money	Millions of $	200.0	24.4	88%	23	22.9	89%	21
Men in Black Money	Millions of $	55.0	39.9	27%	22	26.3	52%	21
The Avengers Critics	Percent	93.0	61.3	34%	24	67.6	27%	21
Men in Black Critics	Percent	67.0	63.9	5%	24	68.9	3%	21
Dow	Percent	-6.2	11.6	286%	24	16.2	361%	21
Apple	Percent	-1.1	17.4	1679%	24	21.2	2026%	21
GE Market Cap	Billions of $	203.7	10.5	95%	24	33.4	84%	21
GM Market Cap	Billions of $	34.8	9.3	73%	24	23.5	33%	21

a. Error expressed as a percentage: Absolute Value((Mean Judgment - True Value) / True Value) * 100

Table 2

Unit of Measurement, True Values, Mean Judgments, and Errors for the Control and Dependent Conditions for Sample Two

			Control			Dependent		
Target Event	Unit	True Value	*M*	Error[a]	*n*	*M*	Error[a]	*n*
U.S. GDP	Percent	2.2	13.5	514%	24	19.3	775%	21
China's GDP	Percent	8.1	29.6	265%	24	34.1	321%	21
Unemployment Rate	Percent	8.1	16.3	101%	24	21.7	168%	21
Retail Sales	Percent	-0.2	6.7	3425%	24	13.9	7055%	21
Gas Price	$	3.9	4.1	7%	24	4.2	8%	21
Gold Price	Hundreds of $	16.6	6.2	62%	24	7.7	54%	21
Home Sales	Millions	5.0	0.1	98%	24	1.4	71%	21
Unemployment Claims	Millions	0.4	3.4	813%	24	4.6	1141%	21
The Avengers Money	Millions of $	200.0	26.8	87%	23	50.2	75%	21
Men in Black Money	Millions of $	55.0	21.5	61%	22	30.4	45%	21
The Avengers Critics	Percent	93.0	59.5	36%	24	56.2	40%	21
Men in Black Critics	Percent	67.0	48.8	27%	24	42.8	36%	21
Dow	Percent	-6.2	15.8	355%	24	2.1	133%	21
Apple	Percent	-1.1	10.7	1072%	24	3.3	395%	21
GE Market Cap	Billions of $	203.7	19.8	90%	24	48.5	76%	21
GM Market Cap	Billions of $	34.8	16.7	52%	24	20.5	41%	21

a. Error expressed as a percentage: Absolute Value((Mean Judgment - True Value) / True Value) *100

Table 3

Collective Error (Deviation from the True Value) for the Highest Ranked, Highest Three Ranked, and All Participants Combined

			Collective Error		
Target Event	Similar Prediction[a]	r^b	Highest	Highest Three	All
U.S. GDP	China's GDP	.55	2.80	**1.47**	17.03
China's GDP	U.S. GDP	.55	**3.9**	4.9	25.45
Unemployment Rate	China's GDP	.34	3.90	**1.90**	8.58
Retail Sales	Unemployment Rate	.22	**10.20**	21.87	14.50
Gold Price	GE Market Cap	.27	15.75	**5.66**	9.47
Home Sales	Dow	.23	4.96	4.60	**4.45**
The Avengers Critics	*Men in Black* Critics	.61	**28.00**	33.67	31.75
Men in Black Critics	*The Avengers* Critics	.61	**5.00**	12.33	10.30
Dow	Apple	.60	**9.20**	9.87	17.59
Apple	Dow	.60	11.10	**7.43**	14.58
GM Market Cap	U.S. GDP	.41	34.81	33.64	**17.63**
Median			9.20	7.43	14.58
Times Most Accurate			5	4	2

Note. Predictions were excluded if no similar prediction was significantly correlated with the target event. Bold values were the most accurate for each prediction.

a. Prediction used to rank participants on accuracy.

b. Correlation between accuracy of the prediction and accuracy on a similar prediction. All correlations are significant at the $p < .05$ level.

Table 4

Results of Significance Tests Comparing Median Individual Error in the Control and Dependent Conditions

	Sample 1				Sample 2			
Target Event	Control Condition Error	Dependent Condition Error	*p*	More accurate	Control Condition Error	Dependent Condition Error	*p*	More accurate
U.S. GDP	19.3	12.8	0.74		2.8	15.8	0.19	
China's GDP	26.9	21.9	0.67		6.9	27.4	0.47	
Unemployment Rate	2.0	5.1	0.72		4.9	3.5	0.84	
Retail Sales	10.2	20.2	0.12		5.2	9.2	0.33	
Gas Price	0.3	0.6	0.01	Control	0.3	0.3	0.59	
Gold Price	13.5	11.6	0.17		15.6	9.2	0.16	
Home Sales	4.9	4.9	0.47		4.9	3.5	0.00	Dep.
Unemployment Claims	0.3	0.2	0.00	Dep.	0.4	4.6	0.02	Control
The Avengers Money	189.0	185.0	0.22		190.0	154.5	0.01	Dep.
Men in Black Money	50.0	38.0	0.03	Dep.	45.0	30.0	0.00	Dep.
The Avengers Critics	23.0	23.0	0.95		28.0	36.5	0.37	
Men in Black Critics	12.5	7.0	0.04	Dep.	13.0	28.0	0.16	
Dow	11.2	21.2	0.05		11.2	8.2	0.01	Dep.
Apple	10.1	21.1	0.07		7.1	4.1	0.00	Dep.
GE Market Cap	204.0	200.0	0.00	Dep.	204.0	150.0	0.01	Dep.
GM Market Cap	34.8	32.3	0.00	Dep.	34.8	19.8	0.02	Dep.

Table 5

Results of Significance Tests Comparing Median Collective Error in the Control and Dependent Conditions

	Sample 1				Sample 2			
Target Event	Control Condition Error	Dependent Condition Error	*p*	More accurate	Control Condition Error	Dependent Condition Error	*p*	More accurate
U.S. GDP	15.3	12.8	0.90		2.7	15.8	0.14	
China's GDP	26.9	21.9	0.73		6.9	27.4	0.32	
Unemployment Rate	1.8	3.9	0.59		4.9	2.9	0.89	
Retail Sales	10.2	20.2	0.02	Control	2.5	9.2	0.02	Control
Gas Price	0.2	0.6	0.00	Control	0.2	0.3	0.47	
Gold Price	12.8	11.6	0.42		15.6	9.2	0.17	
Home Sales	4.9	4.9	0.47		4.9	3.5	0.00	Dep.
Unemployment Claims	0.3	0.2	0.39		0.2	4.6	0.02	
The Avengers Money	190.0	185.0	0.22		191.0	154.5	0.00	Dep.
Men in Black Money	43.7	37.0	0.18		43.0	27.5	0.01	Dep.
The Avengers Critics	23.0	23.0	0.95		28.0	36.5	0.37	
Men in Black Critics	3.0	3.0	0.78		7.0	28.0	0.39	
Dow	11.2	21.2	0.05		11.6	8.2	0.00	Dep.
Apple	10.1	21.1	0.07		7.1	4.1	0.00	Dep.
GE Market Cap	203.6	199.7	0.00	Dep.	203.7	149.7	0.01	Dep.
GM Market Cap	34.8	32.0	0.00	Dep.	34.8	17.3	0.02	Dep.

Table 6

Aggregate Mean Judgments as a Function of Time for Sample One

Target Event	True score	Time period[a] 1	2	3	4	5	6	r^b
				Control condition				
U.S. GDP	2.2	12.8	17.7	19.9	20.4	22.2	20.3	0.83*
China's GDP	8.1	18.8	30.0	29.2	30.8	35.5	37.2	0.91*
Unemployment Rate	8.1	10.5	9.4	11.2	11.5	11.9	16.0	0.84*
Retail Sales	-0.2	9.0	16.6	21.1	17.4	17.2	13.8	0.29
Gas Price	3.9	3.9	3.9	4.0	4.0	4.0	4.0	0.91*
Gold Price	16.6	4.3	2.6	6.7	6.9	7.2	7.2	0.81*
Home Sales	5.0	0.2	0.3	0.7	0.6	0.5	0.4	0.48
Unemployment Claims	0.4	0.5	1.0	0.7	2.9	3.1	2.7	0.86*
The Avengers Money	200.0	7.5	10.3	13.1	22.1	27.4	24.5	0.94*
Men in Black Money	55.0	14.3	12.8	46.7	50.1	46.9	39.9	0.73*
The Avengers Critics	93.0	79.8	68.0	55.6	63.2	62.9	61.3	-0.65
Men in Black Critics	67.0	81.3	77.6	64.5	64.3	66.6	63.9	-0.83*
Dow	-6.2	15.8	12.1	9.6	7.8	11.7	11.6	-0.48
Apple	-1.1	12.0	13.0	12.0	11.9	17.0	17.4	0.80*
GE Market Cap	203.7	1.4	0.8	18.7	15.7	12.6	10.5	0.56
GM Market Cap	34.8	1.0	0.5	18.6	13.8	11.1	9.3	0.51
				Dependent Condition				
U.S. GDP	2.2	11.3	20.8	19.9	25.8	24.0	22.9	0.77*
China's GDP	8.1	40.3	39.5	36.1	32.5	32.2	32.5	-0.93*
Unemployment Rate	8.1	12.0	14.1	13.1	13.4	13.4	13.0	0.24
Retail Sales	-0.2	31.5	29.1	27.3	23.7	22.7	22.2	-0.97*
Gas Price	3.9	5.1	4.7	4.6	4.6	4.5	4.5	-0.84*
Gold Price	16.6	4.7	4.8	5.9	6.5	6.4	7.1	0.96*
Home Sales	5.0	0.0	0.1	0.1	0.1	0.1	0.1	0.72
Unemployment Claims	0.4	0.2	0.2	0.2	0.2	0.2	0.2	0.83*
The Avengers Money	200.0	18.8	19.3	18.2	17.3	16.7	22.9	0.30
Men in Black Money	55.0	13.8	15.3	15.5	17.9	17.9	26.3	0.87*
The Avengers Critics	93.0	67.0	68.4	66.7	67.6	66.9	67.6	-0.03
Men in Black Critics	67.0	78.3	74.0	73.4	71.5	70.8	68.9	-0.96*
Dow	-6.2	19.0	19.0	17.2	17.6	16.4	16.2	-0.93*
Apple	-1.1	25.0	23.8	21.5	20.4	20.0	21.2	-0.85*
GE Market Cap	203.7	2.5	2.1	3.9	17.1	16.3	33.4	0.92*
GM Market Cap	34.8	2.5	2.4	2.4	22.3	22.8	23.5	0.89*

a. Aggregate mean judgments include the noted time period and all previous time periods combined.
b. Correlation between time period and mean judgment.
* indicates $p < .05$.

Table 7

Aggregate Mean Judgments as a Function of Time for Sample Two

Target Event	True score	Time period[a] 1	2	3	4	5	r^b
				Control condition			
U.S. GDP	2.2	6.7	14.0	19.9	14.8	13.5	0.48
China's GDP	8.1	39.3	31.4	42.5	34.4	29.6	-0.48
Unemployment Rate	8.1	20.7	21.7	19.0	17.6	16.3	-0.92*
Retail Sales	-0.2	5.0	0.6	1.2	2.1	6.6	0.29
Gas Price	3.9	4.1	4.1	4.2	4.1	4.1	-0.56
Gold Price	16.6	7.7	6.0	4.4	5.4	6.2	-0.46
Home Sales	5.0	0.1	0.0	0.0	0.1	0.1	0.84*
Unemployment Claims	0.4	1.8	3.0	2.1	3.4	3.4	0.76
The Avengers Money	200.0	6.3	19.0	13.2	17.0	26.8	0.82*
Men in Black Money	55.0	12.0	16.3	12.4	12.7	21.5	0.60
The Avengers Critics	93.0	50.0	52.7	57.9	61.1	59.5	0.92*
Men in Black Critics	67.0	33.3	42.9	46.0	49.7	48.8	0.90*
Dow	-6.2	12.3	7.5	8.6	6.9	15.8	0.27
Apple	-1.1	21.7	13.0	11.3	9.3	10.7	-0.82*
GE Market Cap	203.7	1.3	0.7	0.6	8.1	19.8	0.85*
GM Market Cap	34.8	1.0	0.4	0.3	16.2	16.7	0.86*
				Dependent condition			
U.S. GDP	2.2	26.3	22.3	20.3	18.0	19.3	-0.89*
China's GDP	8.1	42.5	41.3	39.7	35.4	34.1	-0.98*
Unemployment Rate	8.1	30.8	23.5	23.3	23.1	21.8	-0.81*
Retail Sales	-0.2	16.0	12.9	11.1	10.1	13.9	-0.47
Gas Price	3.9	4.0	3.9	4.0	4.2	4.2	0.94*
Gold Price	16.6	7.8	7.2	7.0	8.2	7.7	0.21
Home Sales	5.0	1.6	1.7	1.7	1.5	1.4	-0.59
Unemployment Claims	0.4	4.4	5.2	5.1	4.8	4.6	-0.07
The Avengers Money	200.0	31.0	34.3	37.3	41.8	50.2	0.97*
Men in Black Money	55.0	22.2	26.8	27.6	31.3	30.4	0.92*
The Avengers Critics	93.0	49.0	50.3	51.4	54.3	56.2	0.99*
Men in Black Critics	67.0	44.3	40.6	37.3	38.1	42.8	-0.29
Dow	-6.2	2.3	2.4	1.8	1.8	2.1	-0.51
Apple	-1.1	1.8	2.4	3.0	3.3	3.4	0.95*
GE Market Cap	203.7	57.5	56.7	56.9	51.0	48.5	-0.91*
GM Market Cap	34.8	12.9	27.3	24.6	21.7	20.5	0.28

a. Aggregate mean judgments include the noted time period and all previous time periods combined.

a. Correlation between time period and mean judgment.

* indicates $p < .05$.

Table 8

Results of Significance Tests Between SDs in the Control and Dependent Conditions

	Sample 1			*Sample 2*		
	Control Condition SD	*Dependent Condition SD*	*p*	*Control Condition SD*	*Dependent Condition SD*	*p*
U.S. GDP	26.16	24.70	0.43	19.92	18.07	0.32
China's GDP	25.10	14.54	0.00*	28.68	22.94	0.16
Unemployment Rate	15.11	6.57	0.07	8.40	18.94	0.00*
Retail Sales	24.14	13.16	0.14	22.97	18.23	0.75
Gas Price	0.33	0.89	0.12	0.38	0.41	0.84
Gold Price	7.93	5.14	0.02*	7.26	4.15	0.00*
Home Sales	1.02	0.11	0.02*	0.13	1.13	0.00*
Unemployment Claims	8.03	0.08	0.01*	6.59	3.22	0.04*
The Avengers Money	39.26	30.74	0.20	47.18	33.24	0.44
Men in Black Money	63.94	40.69	0.03*	30.18	19.28	0.42
The Avengers Critics	28.03	12.92	0.00*	22.04	20.98	0.80
Men in Black Critics	25.39	10.23	0.01*	22.63	21.73	0.65
Dow	14.32	12.78	0.55	26.28	2.21	0.00*
Apple	20.05	12.50	0.14	10.26	1.46	0.00*
GE Market Cap	41.38	89.18	0.09	53.51	44.78	0.59
GM Market Cap	40.74	64.43	0.40	49.64	27.09	0.24

Note. For all significant differences the control condition SD is significantly higher than dependent condition SD.

* indicates $p < .05$.

Table 9

Design and Number of Participants in Each Condition in Experiment 2

	No overconfidence information		
	0-trials	2-trials	5-trials
0-cues	27	26	24
2-recent cues	22	26	27
2-long cues	27	21	29
4-cues	28	27	28
	Overconfidence information		
	0-trials	2-trials	5-trials
0-cues	22	18	27
2-recent cues	26	23	35
2-long cues	25	18	19
4-cues	22	24	26

Table 10

Unit of Measurement, True Values, Mean Judgments, and Errors in the Combined Control and Experimental Conditions

			Control			Experimental		
Target Event	Unit	True Value	*M*	Error[a]	*n*	*M*	Error[a]	*n*
Ground Beef Prices	$	3.48	3.94	13%	50	3.45	1%	553
Gas Prices	$	3.86	4.06	5%	50	3.86	0%	553
Unemployment Rate	Percent	7.80	12.19	56%	50	8.45	8%	553
New Unemployment	Millions	1.25	19.67	1477%	50	1.35	9%	553
GDP Change	Percent	2.00	8.22	311%	49	1.84	8%	553
Northeast Home Prices	Thousands of $	1.78	2.61	47%	50	1.65	8%	553
Midwest Home Prices	Thousands of $	1.01	2.11	109%	50	1.01	0%	553
South Home Prices	Thousands of $	1.27	1.84	45%	50	1.34	5%	553
West Home Prices	Thousands of $	1.88	2.57	37%	50	1.73	8%	553
3M Stock Change	Percent	-5.22	8.01	254%	50	1.76	134%	553
Apple Stock Change	Percent	-10.76	14.74	237%	50	4.57	142%	553
GE Stock Change	Percent	-7.27	8.77	221%	50	2.77	138%	553
Microsoft Stock Change	Percent	-4.10	7.36	280%	50	1.11	127%	553
Jackson Touchdowns	Touchdowns/game	0.21	7.37	3337%	50	0.43	100%	553
Turner Touchdowns	Touchdowns/game	0.71	8.44	1082%	50	0.70	2%	552
McGahee Touchdowns	Touchdowns/game	0.40	5.94	1386%	50	0.41	3%	552
Gore Touchdowns	Touchdowns/game	0.57	5.46	855%	50	0.55	4%	548

a. Error expressed as a percentage: Absolute Value((Mean Judgment - True Value) / True Value) * 100

Table 11

Collective Error (Deviation from the True Value) for the Highest Ranked, Highest Three Ranked, and All Participants Combined

		Control				Experimental			
Target Event	Similar Prediction[a]	r^b	Highest	Highest Three	All	r^b	Highest	Highest Three	All
Ground Beef Prices	Unemp. Rate	.34	20.20	5.20	**4.56**	.23	.30	.30	**.28**
Unemployment Rate	GDP Change	.57	2.80	**.72**	11.87	.28	**.30**	.37	.65
GDP Change	Unemp. Rate	.57	**6.20**	9.53	11.65	.28	**6.30**	6.33	6.96
Northeast Home Prices	Midwest Home	.80	**10.75**	13.45	25.54	.39	13.10	**11.97**	15.37
Midwest Home Prices	South Home	.57	.75	**.08**	6.22	.28	.90	.73	**.16**
South Home Prices	West Home	.72	**22.00**	31.30	82.88	.33	39.00	23.00	**13.50**
West Home Prices	South Home	.72	**49.00**	75.70	110.04	.33	5.00	5.30	**.41**
3M Stock Change	South Home	.33	27.00	**3.60**	56.74	.22	**.00**	3.70	6.52
Apple Stock Change	GE Stock	.44	18.00	**4.60**	69.36	.31	**4.00**	6.30	14.56
GE Stock Change	Micro. Stock	.88	**7.20**	11.60	16.07	.16	**1.40**	1.43	10.04
Microsoft Stock Change	GE Stock	.88	**4.00**	8.00	11.46	.16	10.70	**3.47**	10.07
Jackson Touchdowns	Turner T.D.s	.89	**.29**	.32	7.15	.31	**.01**	.22	.21
Turner Touchdowns	McGahee T.D.s	.89	.09	**.04**	7.73	.41	.29	.16	**.01**
McGahee Touchdowns	Gore T.D.s	.61	.02	**.01**	5.54	.42	.20	.15	**.01**
Gore Touchdowns	McGahee T.D.s	.61	**.03**	.08	4.88	.42	.07	.04	**.02**
Median			4.00	3.60	11.46		.90	1.43	.41
Number Most Accurate			8	6	1[c]		6	2	7[c]

Note. Predictions were excluded if no similar prediction was significantly correlated with the target event. Bold values were the most accurate for each prediction.

a. Prediction used to rank participants on accuracy.

b. Correlation between accuracy of the prediction and accuracy on a similar prediction. All correlations are significant at the $p < .05$ level.

c. The number of times the "all" category was most accurate was significantly greater in the experimental than the control conditions, $p = .04$.

Table 12

Individual Error (Deviation from the True Value) and SDs for Judgments from the Combined Experimental Conditions and Combined Control Conditions

	Control			Experimental		
Target Event	*n*	Error	*SD*	*n*	Error	*SD*
Ground Beef Prices	50	1.26	1.13	553	0.14	0.23
Gas Prices	50	0.35	0.41	553	0.14	0.12
Unemployment Rate	50	4.81	9.54	553	0.67	1.10
New Unemployment	50	18.52	49.07	553	0.13	0.30
GDP Change	49	7.09	11.56	553	0.50	0.90
Northeast Home Prices	50	1.08	1.22	553	0.22	0.19
Midwest Home Prices	50	1.21	1.46	553	0.07	0.12
South Home Prices	50	0.85	0.81	553	0.08	0.08
West Home Prices	50	1.09	1.27	553	0.19	0.14
3M Stock Change	50	13.23	13.52	553	6.99	1.64
Apple Stock Change	50	26.27	26.29	553	15.33	3.92
GE Stock Change	50	16.04	11.42	553	10.13	4.43
Microsoft Stock Change	50	12.89	11.81	553	5.63	6.28
Jackson Touchdowns	50	7.16	7.39	553	0.23	0.49
Turner Touchdowns	50	7.81	8.88	552	0.17	0.38
McGahee Touchdowns	50	5.58	7.51	552	0.13	0.32
Gore Touchdowns	50	4.91	7.15	548	0.20	0.65

Note. Error is significantly lower in the experimental groups for all judgments at the $p < .05$ level. SD is significantly higher in control conditions for all judgments at the $p < .05$ level.

Table 13

Correlations Between Condition Mean Judgment Error and Condition SD

Target Event	*r*
Ground Beef Prices	.31
Gas Prices	-.10
Unemployment Rate	.95*
New Unemployment	.95*
GDP Change	-.14
Northeast Home Prices	.10
Midwest Home Prices	.75*
South Home Prices	.09
West Home Prices	-.38
3M Stock Change	.00
Apple Stock Change	-.27
GE Stock Change	-.28
Microsoft Stock Change	-.15
Jackson Touchdowns	.96*
Turner Touchdowns	.84*
McGahee Touchdowns	.94*
Gore Touchdowns	.85*

Note. N = 22.
* indicates $p < .05$.

Table 14

Collective Errors of Means, Medians, and Modes of Judgments

	Control			Experimental		
Target Event	Mean	Median	Mode	Mean	Median	Mode
Ground Beef Prices	0.46	**0.16**	0.48	**0.03**	0.03	0.03
Gas Prices	0.20	0.19	**0.14**	**0.00**	0.01	0.06
Unemployment Rate	4.39	0.30	**0.20**	0.65	0.40	**0.20**
New Unemployment	18.42	4.45	**1.75**	0.11	**0.05**	**0.05**
GDP Change	6.22	1.90	**1.00**	**0.16**	0.20	0.50
Northeast Home Prices	82.88	**59.50**	72.00	**13.50**	18.00	28.00
Midwest Home Prices	110.04	83.50	**49.00**	**0.41**	1.00	1.00
South Home Prices	56.74	33.05	**27.00**	6.52	5.00	**3.00**
West Home Prices	69.36	**42.00**	112.00	14.56	**14.00**	18.00
3M Stock Change	13.23	**8.22**	**8.22**	6.98	6.72	**6.22**
Apple Stock Change	25.50	15.93	**14.76**	**15.33**	15.76	15.76
GE Stock Change	16.04	11.52	**10.27**	**10.04**	10.67	12.27
Microsoft Stock Change	11.46	7.80	**6.10**	**5.21**	6.10	6.10
Jackson Touchdowns	7.15	5.79	**0.79**	0.21	**0.19**	**0.19**
Turner Touchdowns	7.73	5.29	**0.29**	**0.01**	0.06	0.11
McGahee Touchdowns	5.54	2.60	**0.60**	0.01	**0.00**	0.10
Gore Touchdowns	4.88	**1.43**	**1.43**	**0.02**	0.17	0.17
Number most accurate	0	4	12	10	2	3

Note. Bold values are the most accurate in the condition.

Table 15

Standardized Betas for Regression Equations Predicting Judgment Accuracy

Target Event	Knowledge	Similar prediction	Confidence
Ground Beef Prices	0.02	0.32*	-0.06
Gas Prices	0.13*	0.20*	-0.01
Unemployment Rate	-0.03	0.68*	0.03
New Unemployment	-0.04	0.39*	0.02
GDP Change	0.02	0.12*	-0.02
Northeast Home Prices	-0.01	0.02	-0.02
Midwest Home Prices	0.00	0.54*	0.02
South Home Prices	-0.02	0.39*	0.09
West Home Prices	0.00	0.31*	0.04
3M Stock Change	-0.15*	-0.05	-0.04
Apple Stock Change	0.00	0.11*	-0.11*
GE Stock Change	0.05	0.11*	-0.05
Microsoft Stock Change	0.00	0.11*	0.01
Jackson Touchdowns	0.07	0.18*	0.04
Turner Touchdowns	-0.02	0.43*	0.01
McGahee Touchdowns	0.02	0.08	-0.03
Gore Touchdowns	-0.01	0.19*	0.06

Note. Each event was predicted by knowledge, similar prediction, and confidence in a single regression equation.

* indicates $p < .05$.

Table 16

Mean (SD, N) standardized judgment accuracy scores by condition

	No overconfidence information			
	0-trials	2-trials	5-trials	Total
0-cues	-1.72 (1.39, 27)	0.11 (0.20, 26)	0.03 (0.27, 24)	-0.56 (1.20, 77)
2-recent cues	0.10 (0.07, 22)	0.17 (0.19, 26)	0.16 (0.08, 27)	0.15 (0.13, 75)
2-long cues	0.08 (0.15, 27)	0.20 (0.07, 21)	0.18 (0.08, 29)	0.15 (0.12, 77)
4-cues	0.13 (0.06, 28)	0.20 (0.05, 27)	0.15 (0.06, 28)	0.16 (0.06, 83)
Total	-0.37 (1.07, 104)	0.17 (0.15, 100)	0.13 (0.15, 108)	-0.02 (0.67, 312)
	Overconfidence information			
	0-trials	2-trials	5-trials	Total
0-cues	-1.19 (0.67, 22)	0.15 (0.14, 18)	0.09 (0.20, 27)	-0.31 (0.74, 67)
2-recent cues	0.12 (0.05, 26)	0.17 (0.07, 23)	0.18 (0.09, 35)	0.16 (0.08, 84)
2-long cues	0.03 (0.19, 25)	0.20 (0.06, 18)	0.20 (0.09, 19)	0.13 (0.16, 62)
4-cues	0.14 (0.08, 22)	0.19 (0.06, 24)	0.19 (0.12, 26)	0.17 (0.09, 72)
Total	-0.20 (0.64, 95)	0.18 (0.08, 83)	0.16 (0.14, 107)	0.05 (0.42, 285)
	Total			
	0-trials	2-trials	5-trials	Total
0-cues	-1.48 (1.15, 49)	0.13 (0.17, 44)	0.06 (0.23, 51)	-0.44 (1.02, 144)z
2-recent cues	0.11 (0.06, 48)	0.17 (0.14, 49)	0.17 (0.08, 62)	0.15 (0.10, 159)x
2-long cues	0.05 (0.17, 52)	0.20 (0.07, 39)	0.19 (0.09, 48)	0.14 (0.14, 139)x
4-cues	0.13 (0.07, 50)	0.19 (0.06, 51)	0.17 (0.09, 54)	0.17 (0.08, 155)x
Total	-0.29 (0.89, 199)a	0.17 (0.12, 183)b	0.15 (0.15, 215)c	0.01 (0.57, 597)

Note. Means in each row or column that do not share subscripts differ significantly at the $p < .05$ level.

Table 17

ANOVA on Accuracy: 3 (trials: 0, 2, 5) x 4 (cues: zero, 2 recent, 2 long, 4) x 2 (overconfidence information, no overconfidence information)

Source	*Df*	*F*	*η2*	*p*
Trials	2	69.05	0.19	< .001
Cues	3	41.40	0.18	< .001
Overconfidence Information	1	4.39	0.01	0.04
Trials x Cues	6	18.09	0.16	< .001
Trials x Overconfidence Information	2	1.43	0.01	0.24
Cues x Overconfidence Information	3	1.64	0.01	0.18
Trials x Cues x Overconfidence Information	6	0.78	0.01	0.58
Error	573			

Appendix A

Experiment 1 sample procedure

Please write the three digit code that you were emailed.

Code:

We are going to ask you to make a series of predictions. Please use all of your personal knowledge, intuitions, and reasoning ability to make the predictions. Please do not look up additional information as you are making predictions. You will be provided with the average response from other participants who have already made these same predictions. Please feel free to use that information if you wish. It's up to you.

When asked for numbers please write only numbers and not symbols like % or $. Feel free to use decimals or not. You can input positive numbers (like 5.0) or negative numbers (like -5.0).

When reported on April 27, what do you think the percentage growth in real GDP of the U.S. over last year will be? (Real GDP is basically the total economic output of the U.S., a major indicator of how the economy is performing.) The average of previous participants in this study was 18.0 based on 16 participants.

percentage, 0.0 - 100.0:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

When reported for the first quarter of 2012, what do you think the percentage growth in real GDP in China over last year will be? (Real GDP is basically the total economic output of China, a major indicator of how the economy is performing.) The average of previous participants in this study was 35.5 based on 16 participants.

percentage, 0.0 - 100.0:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

When reported at the beginning of this May what do you think the unemployment rate will be, in terms of percent? The average of previous participants in this study was 23.0 based on 16 participants.

percentage, 0.0 - 100.0:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

By what percentage will total retail sales in the U.S. change from the month of April to May (type in a positive or negative number). The average of previous participants in this study was 10.1 based on 16 participants.

percentage, 0.0 - 100.0:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

When measured early this May, what do you think the price of gasoline will be per gallon on average for the U.S.? The average of previous participants in this study was 4.15 based on 16 participants.

Price in dollars and cents: example: 0.00

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

On May 1, what will be the price of one ounce of gold? The average of previous participants in this study was 815 based on 16 participants.

Price in dollars and cents: example: 0.00

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

How many homes will be sold in the U.S. for the month of April? The average of previous participants in this study was 1,484,475 based on 16 participants.

total homes:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

How many individuals will claim unemployment in the U.S. in April? The average of previous participants in this study was 4,796,093 based on 16 participants.

total individuals:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

How much money, in millions, will the movie "The Avengers" make in its opening weekend (Friday, Saturday, and Sunday)? The average of previous participants in this study was 41.8 based on 16 participants.

Millions

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

How much money, in millions, will the movie "Men In Black 3" make in its opening weekend (Friday, Saturday, and Sunday)? The average of previous participants in this study was 31.3 based on 16 participants.

Millions

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

What will be the rating that top critics on the website "Rotten Tomatoes" give the movie "The Avengers?" The average of previous participants in this study was 54.3 based on 16 participants.

Percentage positive (0 to 100)

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

What will be the rating that top critics on the website "Rotten Tomatoes" give the movie "Men in Black 3?" The average of previous participants in this study was 38.1 based on 16 participants.

Percentage positive (0 to 100)

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

For the month of May, what will be the percentage change of the Dow Jones Industrial Average (DJIA, the main stock market)? The average of previous participants in this study was 1.7 based on 16 participants.

% change:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

For the month of May, what will be the percentage change of the computer company "Apple's" stock. The average of previous participants in this study was 3.2 based on 16 participants.

% change:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

What will be the market capitalization (total value of all stock) of General Electric on May 1? The average of previous participants in this study was 51,008,103,125 based on 16 participants.

Total Value:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

What will be the market capitalization (total value of all stock) of General Motors on May 1? The average of previous participants in this study was 21,676,542,847 based on 16 participants.

Total value:

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

Individual difference measures:

What is your gender?

- ❍ Female
- ❍ Male

What is your major?

Taking the good with the bad, how happy and contented are you on the average now, compared with other people?

- ❍ the lowest 5% of the population
- ❍ the lower 30%
- ❍ the middle 30%
- ❍ the upper 30%
- ❍ the highest 5% of the population

Please rate your usual style of thinking on the following scale

- ❍ Highly Intuitive
- ❍ Somewhat Intuitive
- ❍ Equally Intuitive and Analytical
- ❍ Somewhat Analytical
- ❍ Highly Analytical

Please rate your preferred style of thinking on the following scale

- ❍ Highly Intuitive
- ❍ Somewhat Intuitive
- ❍ Equally Intuitive and Analytical
- ❍ Somewhat Analytical
- ❍ Highly Analytical

Isaiah Berlin classified intellectuals as hedgehogs or foxes. The hedgehog knows one big thing and tries to explain as much as possible within that conceptual framework, whereas the fox

knows many small things and is content to improvise explanations on a case by case basis. I place myself toward the hedgehog or fox end of this scale:

- Hedgehog
-
-
-
- Uncertain
-
-
-
- Fox

I think politics is more cloudlike than clocklike ("cloudlike" meaning inherently unpredictable, "clocklike" meaning perfectly predictable if we have adequate knowledge).

- Clocklike
-
-
-
- Uncertain
-
-
-
- Cloudlike

When considering most conflicts, I can usually see how both sides could be right.

- Completely Disagree
-
-
-
- Uncertain
-
-
-
- Completely Agree

Free markets are the best path to prosperity.

- ❍ Completely Disagree
- ❍
- ❍
- ❍
- ❍ Uncertain
- ❍
- ❍
- ❍
- ❍ Completely Agree

I see an irreversible trend toward global economic interdependence.

- ❍ Completely Disagree
- ❍
- ❍
- ❍
- ❍ Uncertain
- ❍
- ❍
- ❍
- ❍ Completely Agree

I am optimistic about the long-term growth trajectory of the world economy.

- ❍ Completely Disagree
- ❍
- ❍
- ❍
- ❍ Uncertain
- ❍
- ❍
- ❍
- ❍ Completely Agree

What was your score on the SAT verbal (critical reading) section? (Please leave blank if you didn't take this test or can't remember).

Score

What was your score on the SAT mathematics section? (Please leave blank if you didn't take this test or can't remember).

Score

Appendix B

Experiment 2 sample procedure

Please enter the three digit code that you were emailed so that we can give you credit for participating.

In this study you will be asked to make predictions and asked about your confidence in the predictions. The individuals with the most accurate predictions, weighted by confidence, will be awarded a $150 gift certificate. The opinion questions at the end of the study will not figure in to the gift certificate calculations. Two certificates will be awarded. If you win, which gift certificate do you want?

- ❍ Amazon.com
- ❍ Apple
- ❍ UConn Co-op

Please type your predicted values into the small boxes on this page. Please just use numbers, do not use symbols like % or $. You may need to use decimals in some cases.

How much does ground beef cost per pound?

How much does regular unleaded gas cost per gallon?

What was the unemployment rate in percent in August of this year?

How many individuals who were looking for their first job were still unemployed last August (answer in millions)?

What is the average yearly GDP growth over the last 20 years?

How much is the national average home price?

How much did the stock price of General Electric increase per year over the last 20 years?

How many touchdowns did Michael Turner score last season (2011 season)?

This study is going to ask you to make predictions about the value of certain future events (unemployment rate, number of touchdowns thrown by Tom Brady, etc.). You will be given a series of variables that you can use to predict that value (see table below). Your job is to try to use this information, along with any personal knowledge you might have, to try to predict the next value. In some cases all of the variables may be useful, in some cases only a few may be useful, and in some cases there won't be much of a pattern at all. If there is no pattern at all, it may be wise to simply use the long term average (usually shown in the last column), because that may be the best guess. In some cases you may learn more after making a few guesses. As a simple example, below is the U.S. population in millions:

September, 2012 value:	1 month before	2 months before	average of entire previous year	overall average from the previous 20 years
?	314.1	313.9	312.1	272.6

the guess for August 2012 might be around 314.3, because there seems to be a minor upward trend.

Individuals tend to be overconfident in their estimates. One way to avoid overconfidence is to think of reasons why your estimate may not be correct. Please rate your confidence carefully.

Please type your answers into the small boxes on the following pages. Please just use numbers, do not use symbols like % or $. Please do not look up any additional information when making your predictions.

Next page:

For the next group of questions you will be predicting the national average price of ground beef per pound.

Ground beef prices in 2010:

September, 2010 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
?	2.85	2.94	2.86	2.20

Your estimate for the 2010 value

Next page:

The real answer is:

September, 2010 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
2.92	2.85	2.94	2.86	2.20

Next page:

Ground beef prices in 2011:

September, 2011 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
?	3.23	3.27	3.10	2.25

Your estimate for the 2011 value

Next page:

The real answer is:

September, 2011 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
3.11	3.23	3.27	3.10	2.25

Next page:

Future prediction:

Ground beef prices in 2012:

September, 2012 Value	1 month before	2 months before	Average of entire previous year	Overall average from the past 20 years
?	3.45	3.45	3.32	2.32

Your estimate for the 2012 value

What is your percentage of confidence in this prediction? Please enter a number from 0 to 100.

0-100

Individual difference measures:

What is your gender?

- ❍ Female
- ❍ Male

What is your major?

Taking the good with the bad, how happy and contented are you on the average now, compared with other people?

- ❍ the lowest 5% of the population
- ❍ the lower 30%
- ❍ the middle 30%
- ❍ the upper 30%
- ❍ the highest 5% of the population

I tend to agree with Democratic politicians on most issues.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I tend to agree with Republican politicians on most issues.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

The hedgehog knows one big thing and tries to explain as much as possible within that conceptual framework, whereas the fox knows many small things and is content to improvise explanations on a case by case basis. I am like a fox:

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

Free markets are the best path to prosperity.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I think rich people make most of their money by exploiting people.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I am optimistic about the long-term growth trajectory of the world economy.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I am worried that environmental problems will eventually lead to disastrous consequences.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I think that in general the condition of the world improves as time moves on.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I am an optimistic person.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I think that most events in life are determined by chance forces that no one can control.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I see myself as dependable, self-disciplined.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I see myself as disorganized, careless.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I am good at coming up with explanations for why things have occurred.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I am good at predicting things.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I am knowledgeable about professional football.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I am knowledgeable about economics.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

What position does Adrian Peterson play?

- ❍ Wide Receiver
- ❍ Running Back
- ❍ Quaterback
- ❍ Tight End
- ❍ I don't know

Which economist first came up with the idea of stimulating the economy through spending?

- ❍ Hayek
- ❍ Keynes
- ❍ Friedman
- ❍ Reagan
- ❍ I don't know

What was your score on the SAT verbal (critical reading) section? (Please leave blank if you didn't take this test or can't remember).

- ❍ Score

What was your score on the SAT mathematics section? (Please leave blank if you didn't take this test or can't remember).

- ❍ Score

I think the predictions I made in this study were accurate.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

I put a lot of effort into this study.

- ❍ Strongly Disagree
- ❍ Disagree
- ❍ Neither Agree nor Disagree
- ❍ Agree
- ❍ Strongly Agree

Please hit the next button (on the bottom right) to submit your survey.