Are Crowds Wise When Predicting Against Point Spreads? It Depends on How You Ask

Joseph Simmons, Yale University, USA
Leif Nelson, University of California at Berkeley, USA
Jeff Galak, Carnegie Mellon University, USA
Shane Frederick, Yale University, USA

Point spread betting markets are considered an important example of crowd wisdom, because point spreads are accurate and are believed to reflect the “crowd’s” predictions of sporting events. However, a season-long experiment found that a sample of football bettors was systematically biased and performed poorly when predicting which team would win against a point spread. Moreover, the crowd’s biases worsened over time. However, when the crowd was instead asked to predict game outcomes by estimating point differentials, its predictions were unbiased and wiser. Thus, the same “crowd” of bettors can appear wise or unwise, depending on how predictions are elicited.

[to cite]:

[url]:
http://www.acrwebsite.org/volumes/15115/volumes/v37/NA-37

[copyright notice]:
This work is copyrighted by The Association for Consumer Research. For permission to copy or use this work in whole or in part, please contact the Copyright Clearance Center at http://www.copyright.com/.
SPECIAL SESSION SUMMARY
Are Crowds Always Wiser?
Andrew Stephen, INSEAD, France
Leonard Lee, Columbia University, USA

SESSION OVERVIEW

Popularized by James Surowiecki’s book of the same title (Surowiecki 2004), the “wisdom-of-crowds” hypothesis states that the aggregation of information in groups can result in better reasons than any single member of the group could make. Surowiecki argues that disorganized group decisions enjoy the advantages of being faster, more reliable, less subject to political forces that can adversely impact decision quality. The popularity of this hypothesis can be attested by the throngs of applications it has spurred and endorsed, ranging from online prediction markets (e.g. NewsFuture, BetFair) to education reform (e.g. Fullan 2004).

This session brings together three recent papers that propose critical boundary conditions for the validity of the “wisdom-of-crowds” hypothesis. Simmons, Nelson, Galak, and Frederick open the session by demonstrating that point spread betting markets, an important application of this hypothesis, lead to systematically biased predictions of NFL football winnings. Even with the opportunity to learn their prediction errors over time throughout the football season, bettors did not improve due to erroneous attributions. Using a longitudinal experiment, they systematically tested the validity of four competing hypotheses and found that the estimation of point differentials, rather than point spread betting, produced significantly improved predictions.

In the same vein of refining the “wisdom-of-crowds” hypothesis, Soll, Larrick, and Mannes posit that “crowds are wise, but well-chosen small crowds are even wiser.” They distinguish between the two extremes of “aggregating the masses” and “chasing the expert” on the continuum of prediction strategies and propose the middle-ground solution of using smaller crowds (e.g. of five people) to improve predictions. Using a range of methodologies — empirical, behavioral, and analytical — they demonstrate convincingly that smaller crowds can indeed be wiser.

Finally, taking things to the other extreme of the continuum, Lee, Pham, and Stephen argue that even with the sole predictor, trusting one’s feelings can significantly improve one’s predictions of a wide variety of crowd behavior over popular prediction markets — from important political outcomes (i.e. the recent Democratic presidential nomination race between Senators Obama and Clinton), to results in the financial (i.e. Dow Jones Index) and entertainment industries (i.e. movie box-office success).

Together, these three papers present diverse perspectives that converge towards the same conclusion — that while crowds may be reasonably wise at times, there are certain interventions or approaches that can be taken to optimize their collective wisdom. Overall, given the fundamental relevance of these papers’ topics to consumers’ everyday lives, this special topic session should be of great interest not only to marketing researchers and psychologists, but also to anyone who is interested in the factors and strategies that can improve our prediction making.

EXTENDED ABSTRACTS

“Are Crowds Wise When Predicting Against Point Spreads? It Depends on How You Ask”
Joseph Simmons, Yale University, USA
Leif Nelson, University of California at Berkeley, USA
Jeff Galak, Carnegie Mellon University, USA
Shane Frederick, Yale University, USA

The wisdom-of-crowds hypothesis predicts that the judgments of a crowd (as measured by any form of central tendency) will be relatively accurate, even when most of the individuals in the crowd are ignorant and error-prone (Surowiecki 2004). Point spreads are often cited as an important example of the wisdom of crowds, because they are very accurate and are widely believed to reflect the “crowd’s” predictions of upcoming sporting events. However, other research (Simmons & Nelson 2006) shows that bettors are biased in their predictions against point spreads: They bet on “favorites” more than “underdogs” despite the empirical observation that the two bets are equally likely to obtain. This research challenges the notion that point spreads capture crowd sentiment, leaving the “wisdom” of the crowd in question.

To test the wisdom-of-crowds hypothesis, we conducted a season-long (17-week) experimental investigation, in which a geographically diverse sample of enthusiastic NFL football fans wagered more than 20,000 on NFL football games against point spreads that were manipulated to favor the underdog (i.e., point spreads that were increased). We investigated four hypotheses. The first hypothesis constitutes the strong version of the wisdom-of-crowds hypothesis and it predicts that crowds will wisely choose against biased point spreads even when they are not told that the spreads are biased (Surowiecki 2004). Our investigation soundly rejected this hypothesis. When predicting against biased point spreads, the crowd predicted vastly more favorites than underdogs, lost more games (and money) than it won, and performed worse than the vast majority of its individual members.

The second hypothesis constitues the weak version of the wisdom-of-crowds hypothesis, which predicts that crowds will wisely choose against biased point spreads when they are told that the spreads have been increased. Although this warning slightly increased the crowd’s tendency to predict underdogs, the crowd nevertheless predicted more favorites than underdogs, lost more games (and money) than it won, and performed worse than most of its individual members. Thus, this more charitable version of the wisdom-of-crowds hypothesis was also rejected.

We investigated a third hypothesis, which asserted that even if crowds are unwise at the start of the study, they should improve over time, as the crowd’s members accumulate evidence of the inferiority of favorites. This hypothesis was also rejected. Moreover, although crowds did not get more or less accurate over time, their predictions did worsen over time in the sense that they unwisely predicted more favorites as the study progressed. Analyses of participants’ tendencies to attribute prediction outcomes to luck or to skill indicate that participants were more likely to attribute correct favorite (vs. underdog) predictions to skill and to attribute incorrect favorite (vs. underdog) predictions to luck. This attributional pattern may have caused people to “learn” that predict-
ing favorites is wiser than predicting underdogs as the study progressed, despite the fact that favorites lost more than underdogs against the spread.

Finally, despite our failure to find evidence for the wisdom-of-crowds when people were asked to predict against point spreads, we did find that a different method of eliciting the same judgment (asking participants to predict point differentials rather than make choices against point spreads) produced vastly different, and vastly wiser, predictions against the spread. In this case, the crowd predicted vastly more underdogs than favorites, won more games than it lost, and outperformed the majority of its individual members. Thus, the same “crowd” of bettors can appear wise or unwise, depending on how their predictions are elicited.

“When it Comes to Wisdom, Smaller Crowds Are Wiser”
Jack Soll, Duke University, USA
Richard Larrick, Duke University, USA
Al Mannes, Duke University, USA

An aggregate opinion of the masses can be more accurate than the best experts. James Surowiecki popularized this idea in his best-seller The Wisdom of Crowds. The power of aggregation is surprising because it contradicts people’s intuition, when faced with a prediction task, to “chase the expert” by seeking out the one person who knows the most. One source of misapprehension is that many people incorrectly believe that an aggregate mirrors the accuracy of its average input (Larrick & Soll 2006). People also mistrust crowds because they fear being dragged down by the crowd’s worst members. This is a valid concern. Aggregation may entail mixing the opinions of skilled and unskilled judges. Chasing might be better and safer, assuming that true experts can be reliably identified.

If chasing means forgoing the statistical benefits of aggregation, and using the whole crowd leaves one vulnerable to incompetence, might there be a middle-ground that employs the best of each strategy? Therein lies the beauty of small crowds. Small crowds achieve most of the benefits of aggregation (Hogarth 1978), and if well-chosen avoid being dragged down by the worst judges.

We examine the wisdom of small crowds from empirical, behavioral, and analytical perspectives for quantitative judgments. First, we compared the small crowd strategy to chasing and averaging the whole crowd in 37 experimental datasets. The chasing and small crowd strategies involve rank-ordering the judges based on their accuracy in a small sample of observations (e.g., one to ten estimates for which the correct answer is known). In the vast majority of cases, the whole crowd beats the best expert, but the small crowd does just as well or better. We also analyzed a real-world dataset—data from an economic forecasting competition sponsored by the Wall Street Journal. The dataset included semi-annual forecasts of macro-economic variables from about fifty economists representing banks, government, and academia. We replicate the wisdom of crowds, but again show that the small crowd strategy outperforms the whole crowd, even when economists are rank ordered based on performance in just one prior forecasting period.

Next, we discuss two experiments that examine beliefs and behavior. In the first experiment, the WSJ panel of economists was described to participants, who then rank ordered five potential strategies for using the forecasts. Participants clearly preferred averaging the top five economists to both chasing the most accurate economist from the previous period and averaging all fifty forecasts. The second experiment used the forecasts of eleven economists from the WSJ survey as stimuli, and required participants to make forecasts over a series of rounds. In each round participants could review summary statistics of the performance of the eleven economists on the preceding rounds. In the “All or One” condition, participants had to decide whether to go with a single expert on a given round or average all eleven forecasts. In the “Small Crowd” condition, participants could average the opinions of any subset of economists, which included the options of selecting one or averaging all. Participants in “All or One” chased a single expert on the vast majority of rounds, and their performance suffered compared to the alternative of averaging all. In contrast, the “Small Crowd” participants tended to include more than one economist in their subset. They typically averaged the forecasts of two chosen economists, which led them to perform better than the “All or One” participants, although not quite as well as they could have done with a somewhat larger small crowd (we recommend five).

Finally, we conducted simulations to investigate the generality and robustness of the small crowd strategy. To do this, we first categorized environments according to three dimensions, corresponding to dispersion in accuracy across judges, correlation in judges’ forecast errors, and the validity of cues to expertise. Averaging the whole crowd performs well when dispersion and correlation are low, and chasing a single expert is ideal when dispersion is high and the better experts can be reliably identified (Soll & Larrick, in press). We find that a small crowd strategy tends to perform closer to the better of these two pure strategies regardless of the environment, and beats them both in many intermediate environments. An interesting result is that small crowds of size five, when selected based on past performance, are robust in the sense that they tend to perform reasonably well across environments, regardless of the size of the crowd from which they are drawn (assuming crowds of size ten or larger).

The fact that a “top five” strategy consistently performs well is surprising, especially in light of the fact that either pure strategy can perform very poorly in the wrong environment. Equally surprising is the fact that the rank-ordering need not be based on a large sample of available data. We find that samples of two or three estimates lead to good results, and the benefits beyond ten are very small. There is a compelling intuition for this result. When there is high dispersion in expertise the better judges are readily apparent with just a few judgments. In contrast, when there is low dispersion in expertise one would require a large sample to obtain a good rank ordering, but in this case it does not matter much which judges are included. Either way, only a small sample is needed.

Crowds are wise, but well-chosen small crowds are even wiser. People do appreciate the wisdom of small crowds, although our research shows that people tend to select crowds that are too small. Our prescription is to rank-order the judges, and then average the top five. We also discuss extensions to consumer research. For example, in choosing hedonic goods such as movies or a vacation destination, consumers can rely on either crowd ratings (available at web sites such as tripadvisor.com and IMDb.com), or on the advice of an “expert” on one’s own preferences, such as a similar other. The small crowds strategy suggests that an aggregate opinion of a small group of similar others may outperform both these strategies.

“The Emotional Oracle: Predicting Crowd Behavior with Feelings”
Leonard Lee, Columbia University, USA
Michel Tuan Pham, Columbia University, USA
Andrew Stephen, INSEAD, France

On August 28, 2008, Barack Obama became the official Democratic candidate for President of the United States in the 2008 presidential election. While this might have seemed inevitable in hindsight, foreseeing it a few months earlier was difficult, especially considering the span and intensity of the contest between him
and Hillary Clinton. Election outcomes, as well as events such as the success of new products (e.g., movies) and stock market movements all depend on the collective actions of masses of people. In this research, we concentrate on the challenge of predicting how such crowds will behave.

Predictions of mass behavior can be made through two distinct processes: (1) a cognitive, scenario-building process, and (2) a feeling-based process that involves the monitoring of one’s subjective feelings toward the options (Dunning 2007, Loewenstein and O’Donoghue 2004). Thus, in predicting crowd behavior, affect might play a role. Whereas the cognitive system of judgment is analytical and logical, the affective system is more holistic and associative (Epstein and Pacini 1999). Consequently, the affective system fosters a more comprehensive processing of available information, distilling the situation to its gist or essential elements (Stephen & Pham 2008). It may also help predictors better relate to—or put themselves in the shoes of—the people whose behaviors they are forecasting.

We examine how reliance on feelings affects the accuracy of peoples’ predictions of crowd behavior across four studies and three diverse contexts. Each study uses the same subtle procedure to induce different degrees to which participants rely on their feelings in making their predictions (Avnet & Pham 2007, Lee, Amir, & Ariely, in press, Stephen & Pham 2008). Participants were randomly assigned to either a “high-trust-in-feelings” (high-TF) or a “low-trust-in-feelings” (low-TF) condition, and asked to describe either two (high-TF) or ten (low-TF) past situations in which they trusted their feelings to make a decision and it emerged as the right decision. Participants typically find it easy to recall two examples of successful reliance on feelings and difficult to recall ten. Thus, those asked to recall two (ten) examples tend to believe that such examples are common (uncommon), and increase (decrease) their reliance on feelings when making subsequent decisions (Avnet & Pham 2007).

In study 1 (N=68; undergraduates), after completing the manipulation, we gave participants information about five movies to be released nationally in early October 2008 (the study was conducted three days before these movies’ release). The task involved ranking these movies in order of predicted success, measured by opening weekend box office revenues. Using the rank-order correlation between each participant’s predicted order and the actual order as a measure of prediction accuracy, high-TF participants were more accurate than their low-TF counterparts. This result held after controlling for prior knowledge and liking of the movies.

Study 2 (N=41; undergraduates) used the same procedure as study 1 and replicated this result (although here we used a set of four movies that were released in mid-December 2008). Additionally, we had participants list all the things that they thought of when making their predictions. We counted the number of items listed that were about mass behavior (and what “other people” would do). As process evidence, we found that a greater percentage of the items listed by high-TF participants were other/crowd-focused and that this was positively correlated with their prediction accuracy. Mediation analysis suggested that the positive effect of reliance on feelings on prediction accuracy was mediated by a greater focus on projections of how “other people” would feel about the movies.

In the third study (N=52; undergraduates), we asked participants (who first completed the trust-infeelings manipulation) to predict the closing values of the Dow Jones stock market index one week in the future (this study was conducted in February 2009). The same basic result from studies 1 and 2 was replicated in this different context: high-TF participants made more accurate predictions. The dependent variable here was the prediction error (actual future value–predicted value); thus, the mean prediction error was lower for high-TF than for low-TF participants. However, we found that this effect only held for participants who possessed some expertise or knowledge of the prediction target (i.e., the economy and stock market): trust in feelings interacted with target knowledge (high for participants who were economics or finance majors, low otherwise) such that high-TF participants were only more accurate if they were knowledgeable.

Finally, in study 4 (N=229; national sample of registered voters, run in mid-February 2008), participants first completed the trust-infeelings manipulation and then predicted the winner of the 2008 Democratic primary. We deliberately ran this study when the Democratic primary contest was far from being conclusive, with both candidates virtually tied in national opinion polls. The results revealed that six months before the Democratic National Convention, high-TF participants were more likely to correctly predict that Obama would win the nomination than low-TF participants. This result held in general, as well as separately among registered Democrats, registered Republicans, and even among participants who had already voted for Clinton.

Overall, we find that feeling-based predictions may lead to greater predictive accuracy of mass crowd behavior. This result holds across different contexts and for short- and long-range predictions. Further, regardless of one’s inherent preference for the collective outcome (e.g., personally liking a particular movie or having already voted for Clinton) this result holds despite cases where the predictor’s personal preference does not align with the crowd. Our results are consistent with the notion that reliance on feelings help people be less encumbered with their own personal preferences or tastes and thus better able to put themselves in the shoes of others—the crowd—to consider how others would feel and what they would do. Accordingly, this leads to more accurate predictions and the counterintuitive finding that focusing less on logic and reasoning and more on feelings enables people to better foresee the future.

REFERENCES

Avnet, T. and M. T. Pham (2007), Metacognitive and Nonmetacognitive Reliance on Affect as Information in Judgment, Yeshiva University, NY.


