Testing the Ability of the Surprisingly Popular Algorithm to Predict the 2017 NBA Playoffs

Michael D. Lee, Julie Vi, Irina Danileiko
Department of Cognitive Sciences
University of California, Irvine

Abstract

We consider the recently-developed "surprisingly popular" (SP) algorithm for aggregating decisions across a group of people (Prelec, Seung, & McCoy, 2017). The algorithm has shown impressive performance in a range of decision-making situations, but has never been tested in terms of making genuine predictions (i.e., where there is no true answer at the time people are asked to make decisions). As a first evaluation of its predictive abilities, we tested the SP algorithm to predict the winners of the 15 match-ups in the 2017 US National Basketball Association playoffs. Our application of the SP algorithm is based on behavioral data—predictions of the winning team, and a meta-cognitive estimate of the percentage of people expected to agree with the predictions—from 100 Amazon Mechanical Turk participants. We compare the accuracy of the SP algorithm to a confidence-weighting algorithm, and to the simple majority rule for a number of data sources. We find that all of the approaches make accurate predictions, typically mis-predicting only one match-up in the playoffs. We discuss the merits of the SP algorithm, and suggest possible future evaluations and analyses.

Keywords: surprisingly popular algorithm, wisdom the the crowd, sporting predictions, expertise, majority rule

Introduction

Prelec et al. (2017) recently proposed the "Surprisingly Popular" (SP) algorithm for aggregating multiple-choice decisions over a group of people. The algorithm is motivated by the challenge of finding accurate answers to single questions, especially in the situation where many people in the group could believe in the wrong answer. For example, if asked whether Chicago is the capital city of Illinois, many people might mistakenly answer "yes". The key feature of the algorithm is that, as well as providing their answer, people are asked to estimate what percentage of other people they expect will also give the same answer. Thus, if somebody knows the capital of Illinois is Springfield, but also realizes that most people mistakenly believe it is Chicago, both of those pieces of information can be expressed. The knowledgeable person says "no" to Chicago, but then says few people will agree with them.

The SP algorithm combines the cognitive judgment (the basic decision) and the meta-cognitive judgment (the estimate of the decisions of others) by comparing the expected and observed proportion of people in a group making a decision. The observed proportion is simply how many say "yes" to Chicago being the capital. The expected proportion combines the estimated percentages for those who say both "yes" and "no". A person who says "yes" and expects 80% of people to agree contributes to the expected proportion in the same way as person who says "no" but expects only 20% to agree. The final decision made by the SP algorithm compares the observed and expected proportions, and chooses the answer that has more observed agreement than is expected (i.e., the answer that is "surprisingly popular"). Intuitively, people who believe Chicago is the capital of Illinois will tend to believe others will say the same, while those who believe it is not expect others to disagree. Thus, the expected agreement is very high, so is the observed agreement is lower—because some knowledgeable people say "no"—then "no" will be the surprisingly popular answer.

Prelec et al. (2017) evaluate the accuracy of the SP algorithm in a number of domains, including trivia and general knowledge questions, medical diagnoses, and art price category evaluations. For all of these domains, the SP algorithm achieves impressive levels of accuracy, outperforming standard alternatives like the majority answer, and the answer in which people express the greatest overall confidence. While the trivia and general knowledge domains have limited real-world applicability—outside the confines of a trivia competition, it is possible simply to look up the answers—the medical diagnosis and art evaluation domains clearly could have useful application.

In both cases, an accurate decision based on cheap and simple behavioral judgments is a useful capability, since determining the "true" answer involves expensive medical testing in the first case, and time-consuming and complicated mechanisms like art auctions in the second case.

Perhaps the most interesting and important potential application of the SP algorithm, however, is to genuine predictions. These are forecasting situations where the true answer is not known at the time people make decisions. None of the domains considered by Prelec et al. (2017) are of this type. The goal of the current research is to provide a first evaluation of the SP algorithm in a predictive setting. Our predictive setting is the US National Basketball Association (NBA) 2017 playoffs.

Experiment

The NBA playoffs involve a total of 15 match-ups over 4 rounds in a standard bracket structure. There are 16 teams in the first round, paired into 8 match-ups based on their conference seeding. The winners progress into a second round involving 8 teams, and 4 match-ups. The 4 winners play in a third round of"conference finals" involving 2 match-ups, and the two conference champions are matched-up in a fourth round "finals series".[1] Throughout all four rounds, each match-up involves a best-of-7-game series, so that once one team has won 4 games, the match-up is decided.

Before each round, we collected predictions about each match-up in the round from 100 participants, using the Amazon Mechanical Turk (AMT) system.[2] Participants first provided basic demographic information: their gender, and their age range from the options 18–24, 25–34, 35–44, 45–54, 55–64, and 65+. They then rated their knowledge of NBA basketball on a 5-point scale: extremely knowledgeable, very knowledgeable, moderately knowledgeable, slightly knowledgeable, and not knowledgeable at all.

[1] As always in American sport, the overall winner is declared the "World Champion", despite no teams outside North America participating in the league.

[2] In two cases, the first game of a match-up from a later round was played before the final game of a previous round was completed. In these situations, we collected data once the previous round was complete, and all the teams and match-ups in the current round were known. This meant that, on two occasions, data were collected once one game in the series about which predictions were being made had already been completed. These completed results were a Celtics win over the Wizards on April 30 before the Clippers versus Jazz result was known later that day, and a Warriors win over the Spurs on May 14 before the Celtics versus Wizards outcome was known on May 15.

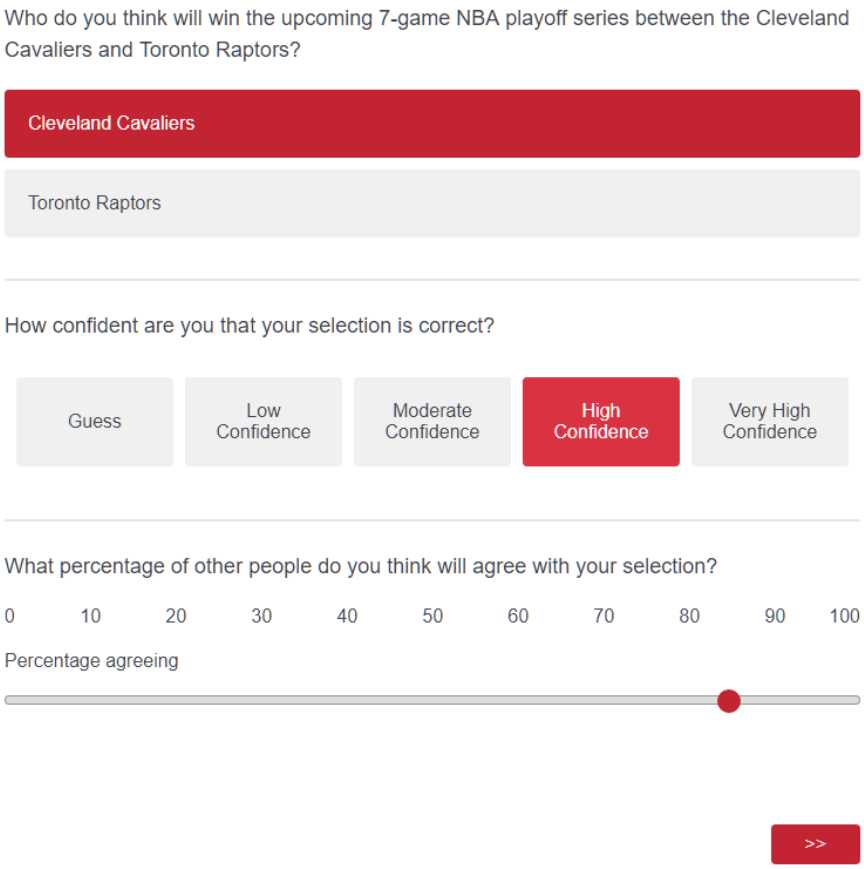

Figure 1. Example of the experimental interface for making predictions about a match-up. participants are first asked to predict the team they believe will win the match-up, then express a confidence in their prediction, before estimating using a slider the percentage of other people they believe will agree with their prediction.

The predictions about the match-ups were completed using the interface shown in Figure 1. First, they made a prediction about which team would win the match-up, then they rated their confidence in this prediction on a 5-points scale, and finally they estimated the percentage of others they believed would agree with their prediction. This estimation was done with a slider that showed the exact whole number from 0 to 100 being selected as the slider was moved, as well as permanent indicators $0, 10, \ldots, 100$ on a scale. Once a participant had completed these three behavioral responses for a match-up, they pressed an "advance" button that presented the same interface for the next match-up in the round. They could adjust the behavioral responses for a specific match-up while it was presented, but could not return to an

earlier match-up. Once all match-ups were completed, the experiment was concluded. Participants were paid US$1 for rounds 1, US$0.50 for round 2, and US$0.25 for rounds 3 and 4, roughly in proportion to the number of games involved in each round. Different AMT workers participated in each of the 4 rounds of data collection.

Results

We first discuss the results for the SP algorithm, then for the confidence-weighting approach, and then finally for a simple majority-rule approach. For the majority-rule approach, we consider some publically available predictions from basketball experts and fans, as well as the predictions we collected from MTurk.

SP Algorithm

The performance of the SP algorithm for all 15 match-ups is shown in Figure 2. Each panel represents a match-up, and has the same structure. The two teams are listed, with the home team listed above the away team. The distribution of meta-cognitive estimates is shown by the bar graphs, with estimates made by people predicting the home team to win shown in the upper (blue) bar graph, and the estimates made by people predicting the away to team to win shown by the lower (gray) bar graph. The distributions range from 0 to 100 from left to right, with each bin having a width of 10 percentage points. The total area in each bar graph is proportional to the number of people who chose the home and away teams. From these proportions, and distributions of estimates, the observed and expected percentage of people predicting the home team will win can be calculated, as shown by the solid line labeled "E" and the dashed line labeled "O" respectively. These percentages are also listed in each panel, along with the prediction made by the SP algorithm, which chooses the team with greater observed than expected support. Finally, the accuracy of the prediction is shown by a tick or cross.

Figure 2 shows that the SP algorithm correctly predicted 14 of the 15 match-ups. The incorrect prediction was in favor of the Clippers over the Jazz in the first round. The distributions of the meta-cognitive estimates show some interpretable regularities. The favored team — that is, the team that the majority of people selected — generally has higher estimates of agreement. An extreme example is the Warriors versus Trail Blazers match-up in the first round, for which many people selected the Warriors, and expected almost everybody else to agree with them. In contrast, there are few

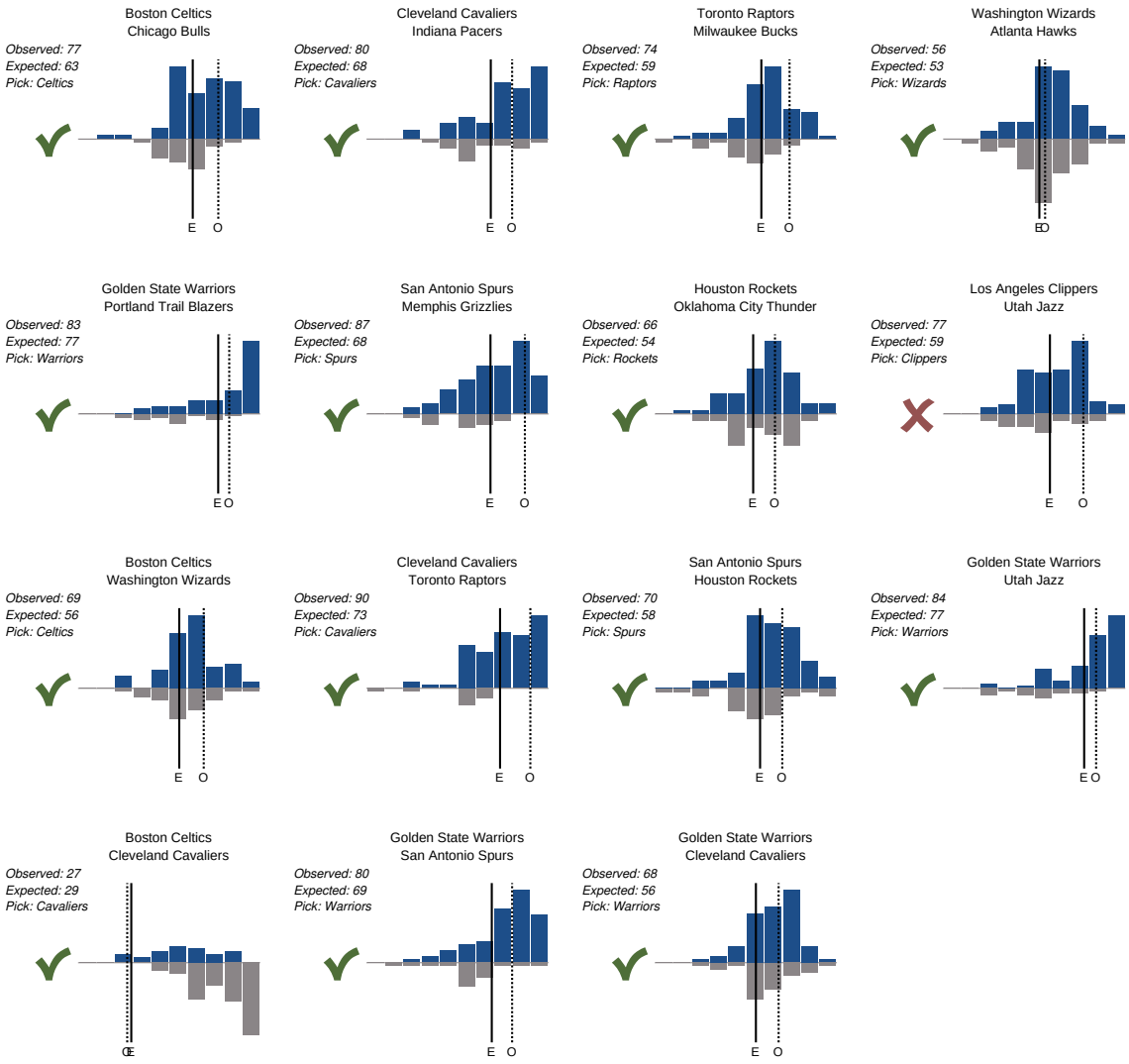

Figure 2. Performance of the SP algorithm. Each panel corresponds to a match-up, with home team predictions in the upper (blue) bar graph, and away team predictions in the lower (gray) bar graph. The distributions show the meta-cognitive estimates of agreement from 0 to 100 made by people selecting each team. The solid "E" line and dashed "O" line show, respectively, expected and observed percentages in favor of the home detail. These percentages and the prediction of the SP algorithm are listed, and the accuracy of the prediction is shown by a tick or cross.

instances of people selecting the less-favored team and expecting high agreement. This suggests that people are aware when they are selecting the underdog, and can express this information through their meta-cognitive judgment of agreement.

Confidence-Weighted Predictions

The performance of a confidence-weighted algorithm is shown in Figure 3. Each panel again corresponds to a match-up. The distributions now show the confidence ratings—on a 5-point scale—for the home and away teams. The solid lines show the mean confidence rating for people favoring each team. The tally of the total confidence in each team—that is, the sum of all the confidence ratings given by people favoring that team—are detailed. The predictions of the confidence-weighting algorithm, which is the team with the greatest total confidence, is also detailed. Figure 2 shows that the confidence-weighting approach leads to the same decisions as the SP algorithm, and so makes the same incorrect predictions for the Clippers versus Jazz match-up.

Majority Predictions

The performance of the majority rule, which simply predicts the team favored by more than 50% of people, is shown in Figure 4. Each panel again corresponds to a match-up, and the bars within each panel correspond to different data sources. The "MTurk" source corresponds to the data we collected, and is simply the team predicted by participants. The "ESPN Expert" source is a collection of expert pundit predictions from the sporting website ESPN.[3] The "CARMELO" source is a prediction made by an algorithm developed by the data analysis website `fivethirtyeight.com`. Unlike the other data sources, it is based on basketball statistics (i.e., measures of team and individual performance) rather than consolidated human opinion. The final "ESPN Crowd" data source, which was only available for the first round of the playoffs, is a large number of visitors to the `espn.com` website who made predictions. Each bar shows the proportion of people favoring the home team for that data source, except for the CARMELO data source, which is a rating number between 0 and 1, These bar heights are compared to the 0.5 proportion line. The accuracy of the resulting prediction is indicated by a tick or a cross above each bar.

[3] An example of this data source is (currently) available at `http://www.espn.com/nba/story/_/id/19279167/2017-nba-playoffs-expert-predictions-conference-semifinal-matchups`

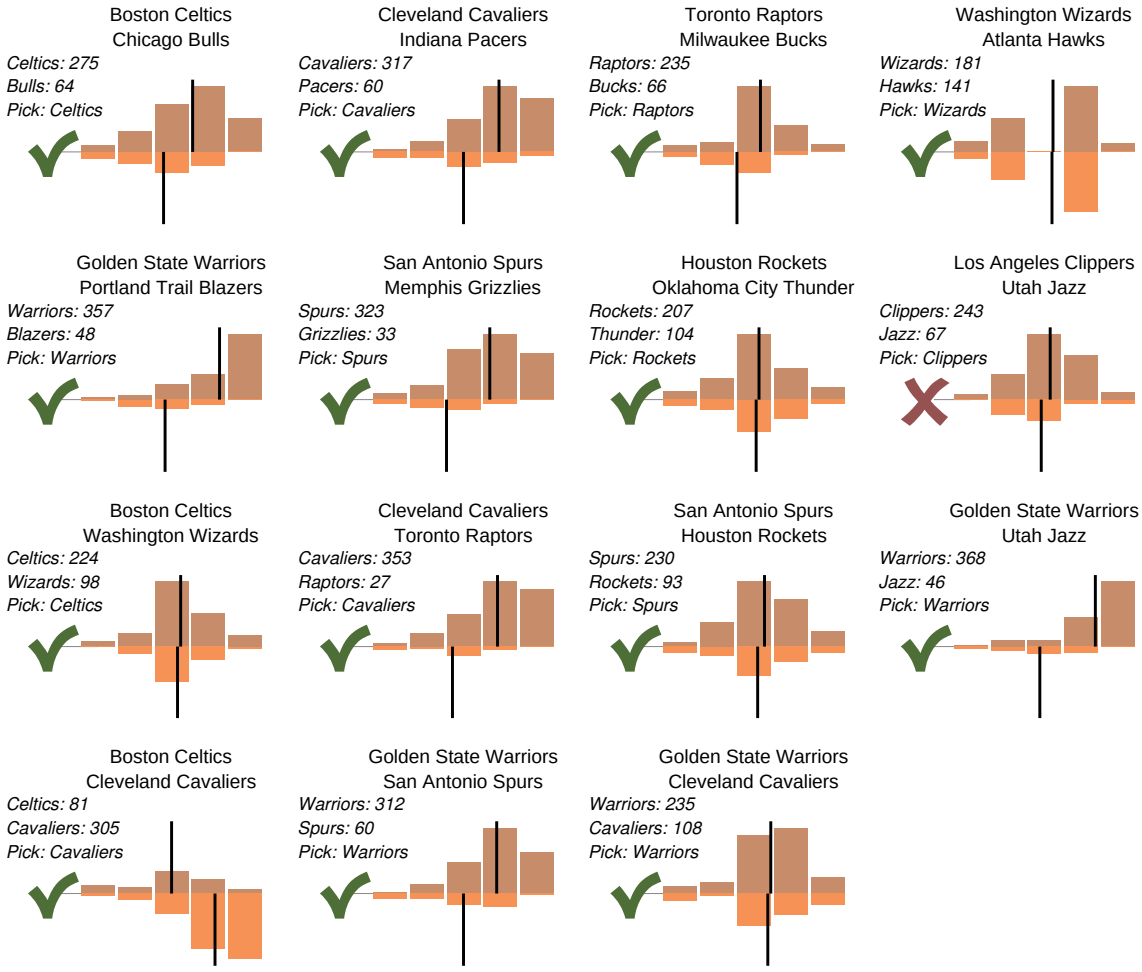

Figure 3. Performance of the confidence-weighting algorithm. Each panel corresponds to a match-up, with home team predictions in the upper (blue) bar graph, and away team predictions in the lower (gray) bar graph. The distributions show the confidence ratings made by people selecting each team. The solid lines show the mean confidence for predictions in favor of each team. The total confidence for each team is detailed, and team predicted by the confidence-weighting algorithm is the one with the maximum confidence total. The accuracy of the prediction for each match-up is shown by a tick or cross.

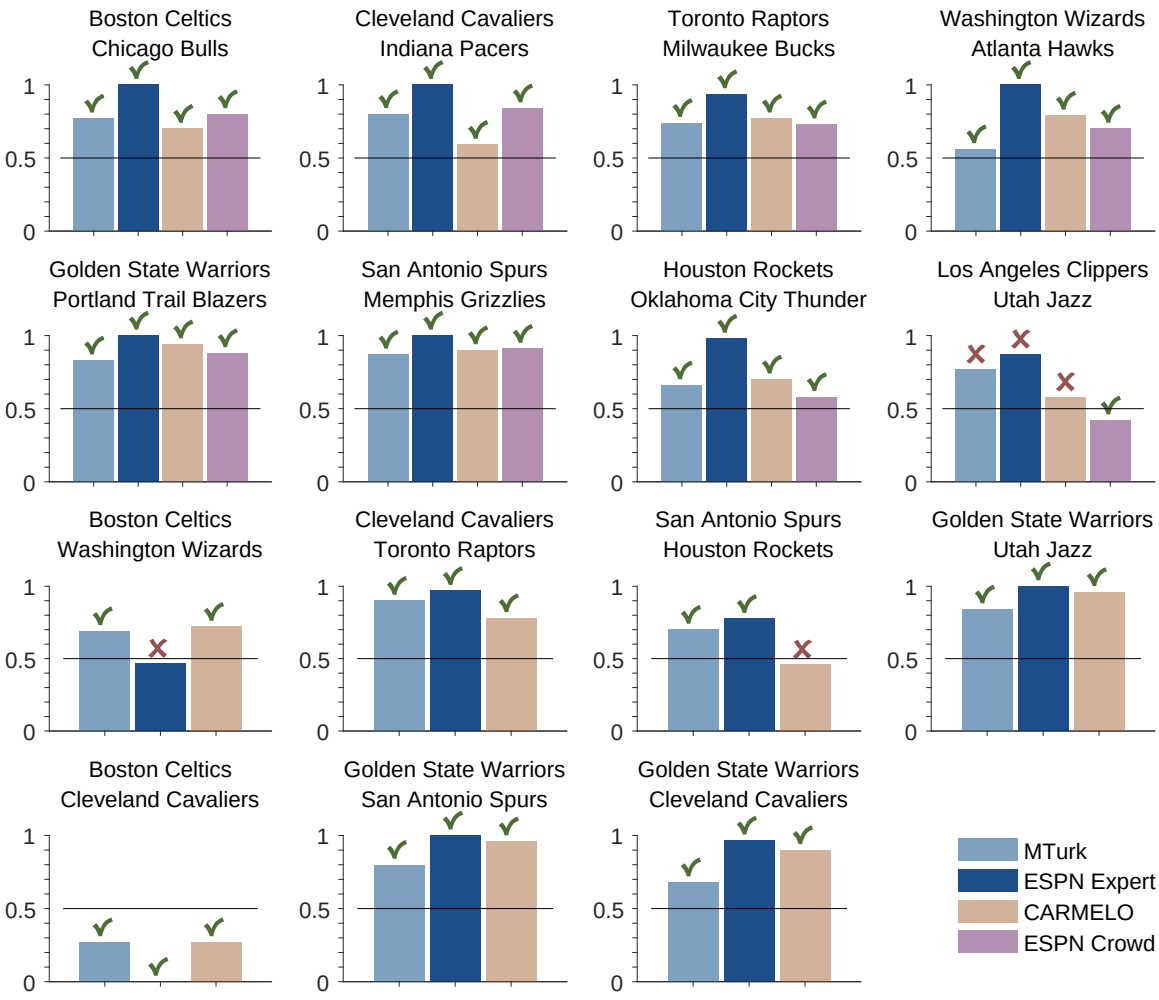

Figure 4. Performance of the majority rule. Each panel corresponds to a match-up. The individual bars correspond to the proportion of people in an empirical data source who predicted the home team. The majority rule predicts the team favored by more than 50% of people, and the accuracy of these predictions are shown by ticks or crosses above each bar. Note that the "ESPN Crowd" data source was only available for the first round of the playoffs, and so it only shown for the first 8 match-ups.

Figure 4 shows that, for most data sources and match-ups, the majority rule makes the same predictions as the SP and confidence-weighting algorithms. There are, however, a few interesting exceptions. The ESPN Crowd correctly predicted the Jazz to defeat the Clippers, ESPN experts narrowly but incorrectly predicted the Wizards to defeat the Celtics, and the CARMELO algorithm incorrectly predicted the Rockets to defeat the Spurs. Overall the MTurk majority makes one incorrect prediction, while the ESPN experts and the CARMELO algorithm make two incorrect predictions. The ESPN Crowd made completely accurate predictions for the 8 match-ups in the first round of the playoffs for which it was available.

Discussion

A total of 15 decisions is not enough to make a definitive evaluation of a group aggregation method, and the 2017 NBA playoffs proved to be particularly uninformative. Most match-ups were won by the widely-favored team. The SP algorithm correctly predicted all of these expected victories, but so did standard algorithm based on confidence-weighted judgments, and simple majorities. The one match-up that might be regarded as an upset — the Jazz victory over the Clippers in the first round — was mis-predicted by all of the approaches we tested that used human judgments as their data source. Only the CARMELO algorithm, an analysis based on performance statistics, correctly predicted the Jazz.

Accordingly, we view the current contribution as a first demonstration of the applicability of the SP algorithm to making predictions, with a particular focus on the important class of predictions represented by sporting contests. It seems clear that MTurk participants were able to make decisions and provide meta-cognitive judgments in a prediction setting as naturally as they are able in previously studied non-prediction settings like general knowledge questions. Our best guess is that, if a subset of people have more insight into a surprise winner in a sporting contest, the SP algorithm provides a simple and effective way to capture and use that knowledge. The question is whether and how often such subsets of people exist.

Beyond the direct evaluation of the accuracy of the SP algorithm, our data and results suggest a number of lines of future research. One involves evaluating the *calibration* of the various algorithms. Ideally, an algorithm would always make correct predictions, but where errors are inevitable, it is important to know how much confidence should be placed in a prediction. A calibration curve describes

the relationship between some output of an algorithm that expresses confidence in a decision, and the actual empirical frequency with which decisions made with that level of confidence are correct. For the SP algorithm, one candidate output is the difference between the expected and observed percentages for a team, since it is their difference that determines the prediction. For the confidence-weighting algorithm, an obvious candidate is the difference between the confidence tallies. For majorities, it is the difference between the proportions in favor of each team. Whether these differences provide regular and useful signals about the accuracy of predictions is an important topic for future research. Lee and Lee (2017) provide an example of the sort of analysis that could be pursued, for the case of the majority. It is not obvious to us that the SP algorithm will have good calibration properties, but this remains an open question in need of more data and systematic evaluation.

Other lines of future research is suggested by the meta-cognitive judgment of agreement that is the novel empirical feature of the SP algorithm. The distributions of estimates of agreement shown throughout Figure 2 are generally broad. Whether this arises because of a individual differences in opinion about the percentage, individual differences in the cognitive processes in the way in the estimates are generated, or both, is an interesting cognitive modeling question. Potentially, a model-based account of these data could be incorporated into the SP algorithm to improve the precision of the meta-cognitive estimates. Ideally, differences due to the cognitive processes involved in meta-cognition could be "factored out", leaving the key signal of expected agreement. More specifically, the meta-cognitive estimate of agreement provides an interesting binary signal when it is judged to be below 50% of others. This estimate means that a person is making what they believe to be a minority or underdog prediction. It would be interesting to study how often and when these estimates are made, and whether the signal they provide can extend or refine the SP algorithm.

Acknowledgments

We thank Drazan Prelec and colleagues for supplying raw data that helped us verify our implementation of the SP algorithm. All of the raw data used in this paper is available at the Open Science Framework project page at `https://osf.io/3kjmu/`.

References

Lee, M. D., & Lee, M. N. (2017). The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making*, *12*, 328–343.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*, 532–535.