Corrigendum for "Testing the Ability of the Surprisingly Popular Method to Predict NFL Games"

Michael D. Lee, Irina Danileiko, Julie Vi

Department of Cognitive Sciences

University of California, Irvine

Abstract

This corrigendum corrects errors in Lee, Danileiko, and Vi (2018). The errors relate to the performance of the surprisingly popular and confidence-weighted tally methods for AMT participants who self-rated as "extremely knowledgeable" about the NFL. The original analyses involved a coding error that led to the wrong meta-cognitive estimates and confidence ratings being used for some analyses, leading to errors in calculating the accuracy of the predictions for these methods. We present corrected versions of all of the analyses affected by this error, and updated discussion in light of the corrected results.

Corrected Analyses

The errors relate only to the performance of the surprisingly popular and confidence-weighted tally methods for AMT participants who self-rated as "extremely knowledgeable." This means a number of the analysis in Lee et al. (2018) are not affected by the error. These include the distribution of accuracy, confidence and meta-cognitive judgment by self-rated expertise in Figure 2, the calibration curve analysis in Figure 6, and other analyses that do not involve the subset of "extremely knowledgeable" AMT participants. The analyses that do need to be corrected are presented below.

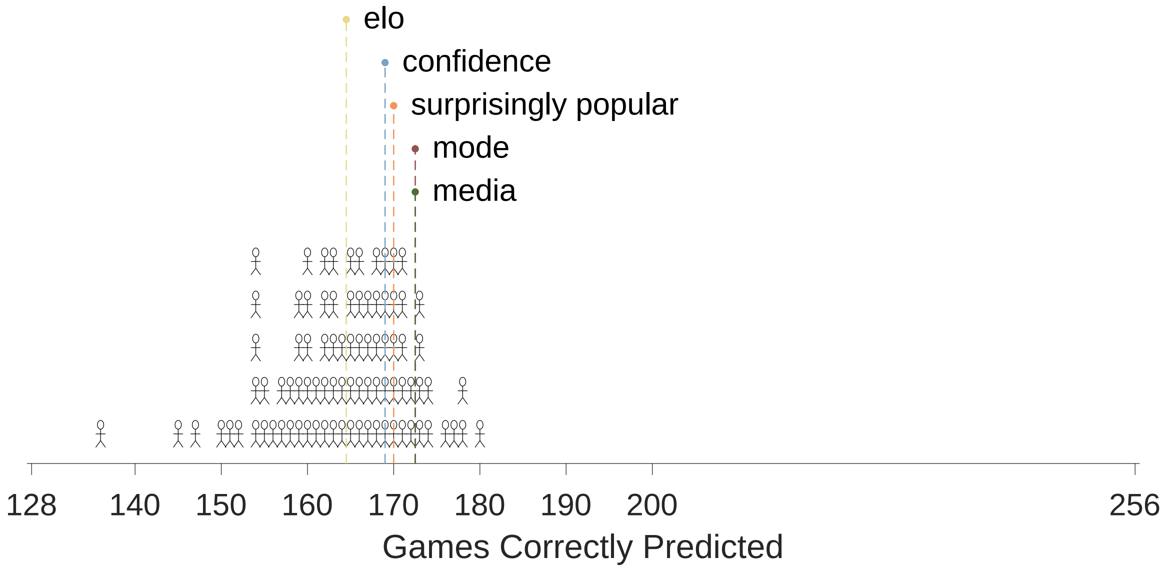

Figure 1. Correction for Lee et al. (2018, Figure 3). The number of games correctly predicted by the surprising popular (sp), confidence-weighted tally (conf), mode of AMT participants (mode), mode of media experts (media), and Elo (elo) methods, and for 94 individual media experts. The stick figures represent the distribution of the number of games correctly predicted by the media experts. The labeled lines show the number of games correctly predicted by the methods.

Accuracy of the Surprisingly Popular Method and Other Predictions

Figure 1 summarizes the overall performance of the surprisingly popular method, in the context of the performance of the individual media experts and other comparison methods. The stick figures show the distribution of the total number of games in the season correctly predicted by the 94 individual media experts. The worst-performed experts correctly predicted the outcome of around 150 of the 256 games, the best-performed experts correctly predicted about 180 games, and most experts were correct for between 155 and 175 of the games.

Once corrected, the surprisingly popular method, based on the group of AMT participants self-rated as "extremely knowledgeable" predicted 170 games (rather than 174 games) correctly. This performance was inferior to 19 of the media experts, the same as six more, and superior to the remaining 69. The surprisingly popular method now slightly underperforms the modal (most common) predictions of the "extremely knowledgeable" AMT participants and the modal predictions of the media experts.

Examples of the Surprisingly Popular Method

Two of the three illustrative examples used by Lee et al. (2018) are better replaced once the analyses are corrected, using different games to make the same conceptual points. Figure 2 presents the revised examples of the performance of the surprisingly popular method, and the confidence-weighted tally method, aimed at giving some insight into the relative success of the surprisingly popular method. The detailed results for every game are available in the supplementary material. The three examples in Figure 2 were chosen because the surprisingly popular and confidence-weighted tally methods make different predictions and because, in two cases, the surprisingly popular prediction does not follow the majority. It is the ability of the surprisingly popular method to predict against confidence-based and majority opinion that makes it theoretically interesting.

In the Browns versus Steelers game, a large majority of participants predict the Steelers to win, and most express high confidence in this outcome and expect other people overwhelmingly to agree with them. For the confidence-weighted tally method, the high-confidence predictions of a Steelers victory lead to it being chosen. For the surprisingly popular method, it is the difference between observed and expected predictions that is important. The overall expectation is that only 20% of people will predict the Browns to win. Thus, when 22% of people do make this prediction, the surprisingly popular method chooses the Browns. As the tick and cross marks in Figure 2 indicate, the winner of the game was the Steelers, and so the surprisingly popular method makes an incorrect prediction. Despite the incorrect prediction, this game provides a clear example of how the mechanics of the surprisingly popular method can lead to a team being chosen that most people did not predict to win.

In the Vikings versus Saints game, the predictions of the participants are evenly split between the two teams. The surprisingly popular method predicts a Vikings victory because overall more participants favoring the Vikings believe relatively few others will agree with them. Thus the observed 50-50 split exceeds the expectation that only 46% of participants will choose the Vikings. Meanwhile, the confidence-weighted tally method predicted a Saints victory, since participants favoring the Saints often expressed high confidence while participants favoring the Vikings generally expressed less confidence in their predictions. As it turned out, the Vikings won the game, so the surprisingly popular method made the correct prediction.

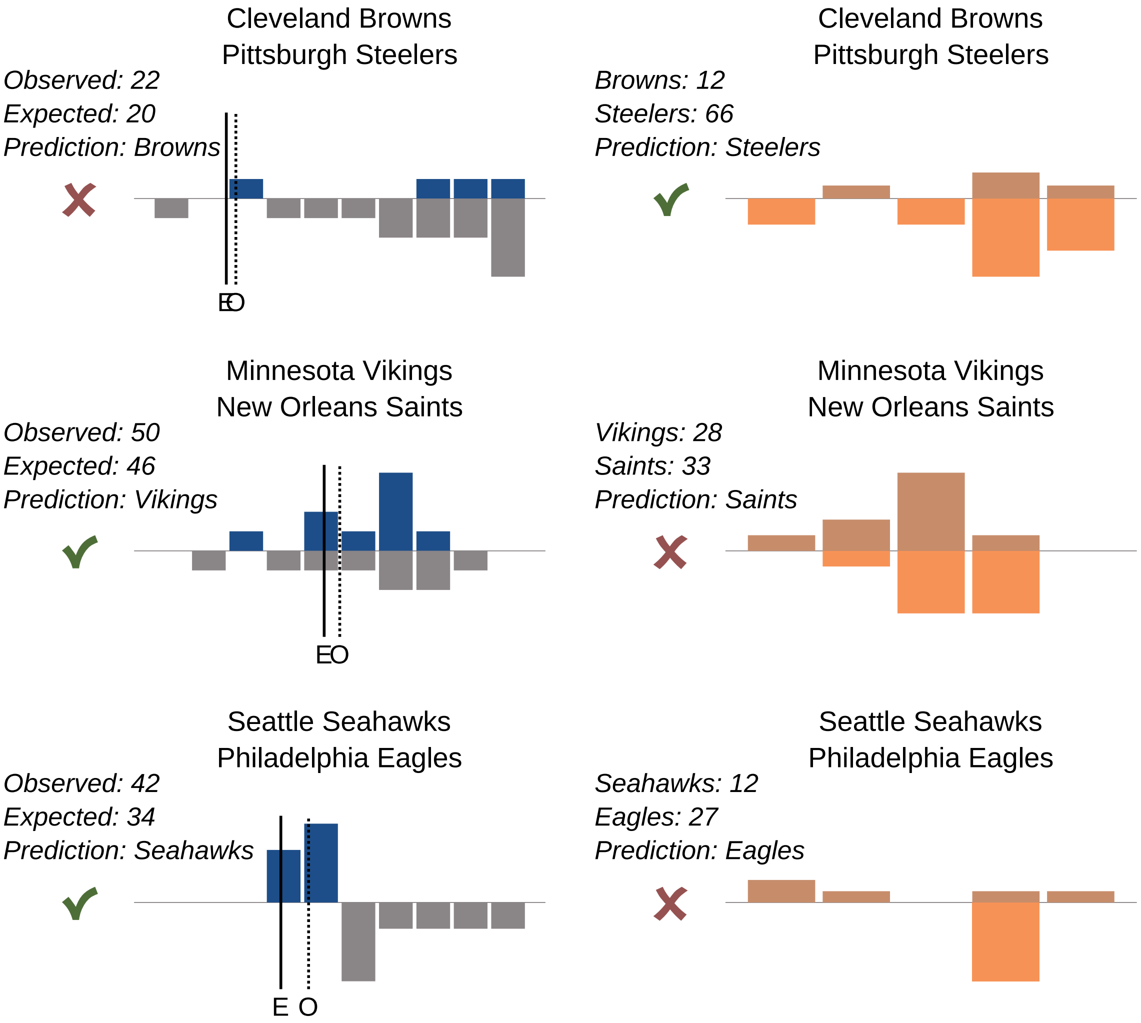

Figure 2. Correction for Lee et al. (2018, Figure 4). Illustrative examples of the surprisingly popular and confidence-weighted tally methods for three NFL games. The three games correspond to the rows of panels, with the left-hand panel corresponding to the surprisingly popular method, and the right-hand panel corresponding to the confidence-weighted tally method. For the surprisingly popular method, the distributions of meta-cognitive estimates of agreement are shown for people choosing each team, and the observed and expected percentages of first-named home-team prediction are detailed, along with the answer of the method and its accuracy. For the confidence-weighted tally method, the distributions of confidence on a 5-point scale are shown for people choosing each team, and the confidence tallies are detailed, along with the answer of the method and its accuracy.

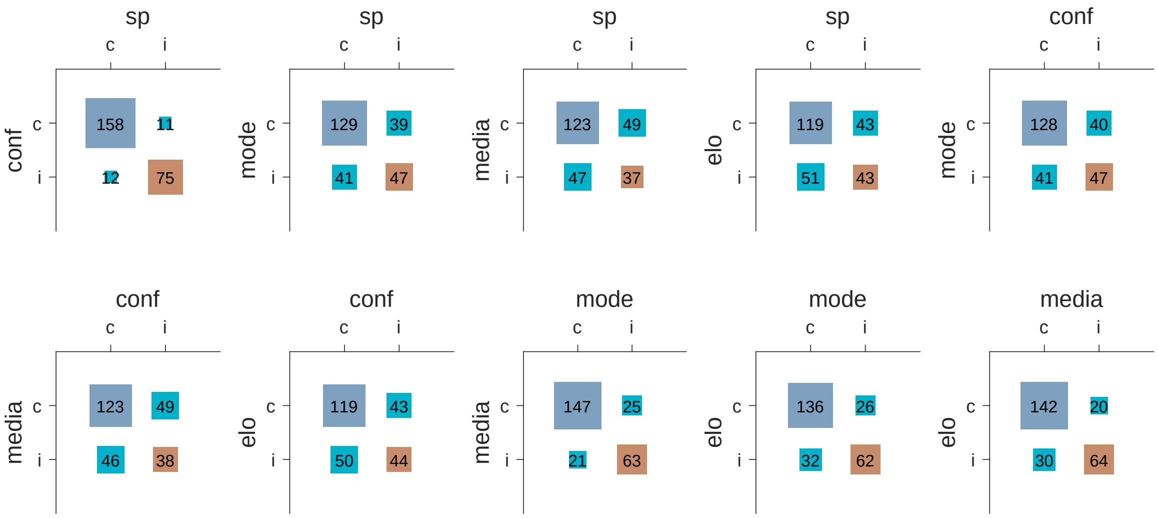

Figure 3. Correction for Lee et al. (2018, Figure 5). Relationship between pairs of methods and the accuracy of their predictions. Each panel corresponds to one of the 10 unique pairings of the five methods: surprising popular (sp), confidence-weighted tally (conf), mode of AMT participants (mode), mode of media experts (media), and Elo (elo). Within each panel, correct predictions are labeled as "c" and incorrect predictions are labeled as "i". The areas of squares and overlain numbers show counts of games in which both methods made the same correct prediction (top-left), the same incorrect prediction (bottom-right), the left-labeled row method made a correct prediction but the top-labeled column method did not (top-right), or the top-labeled column method made a correct prediction but the left-labeled row method did not (bottom-left)

.

Finally, in the Seahawks versus Eagles game, there is an expectation that the Eagles are heavy favorites. Most participants who predict an Eagles win believe the majority of others will agree with them. Most participants predicting a Seahawks win believe the majority will disagree with them, but will instead favor the Eagles. Thus, even though a minority of 42% of participants predicted the Seahawks to win, this is higher than than the 34% expectation, and the surprisingly popular method correctly predicts the actual Seahawks win.

Overlap in Predictions

Figure 3 provides a summary of the relationship between the predictions of the different methods and their accuracies, using corrected analyses for the surprisingly popular and confidence-weighted tally methods. The results in Figure 3 are very similar to those originally presented, and not require a revision to the general findings or conclusions presented in Lee et al. (2018). The exploratory boosting result mentioned by Lee et al. (2018), involving the combined performance from taking the mode of all five methods, becomes 177 rather than 178 games correct.

All Predictions

Figure 4 details the accuracy of every aggregate prediction method for every game. Each panel corresponds to a week, labeled "w1" for week 1, and so on. Within each panel, rows correspond to methods, and columns to games, ordered from left to right from best overall predicted to worst overall predicted. A dark blue circle indicates a correct prediction; a light orange circle indicates an incorrect prediction; a gray circle indicates neither team was favored by the method.

Conclusion

Our corrected results continue suggest there is promise in applying the surprisingly popular method to predicting the outcomes of NFL games, although the correct results show performance slightly below rather than slightly above the real-world benchmarks. We continue, as Lee et al. (2018) said, to

> " ... recognize the limitations of evaluating of any prediction method based on only one season, comprising a few hundred binary outcomes. Accordingly, we view this study as a motivating demonstration of the applicability of the surprisingly popular method to making predictions, with a particular focus on the important class of predictions represented by sporting contests. It seems clear that people 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."

Similarly our overall conclusion remains the same as the original one given by Lee et al. (2018):

> "The key question, therefore, is whether and how often such subsets of people exist. Our study provided some first suggestive evidence that they can exist and suggests that domain knowledge, as measured by self-rating

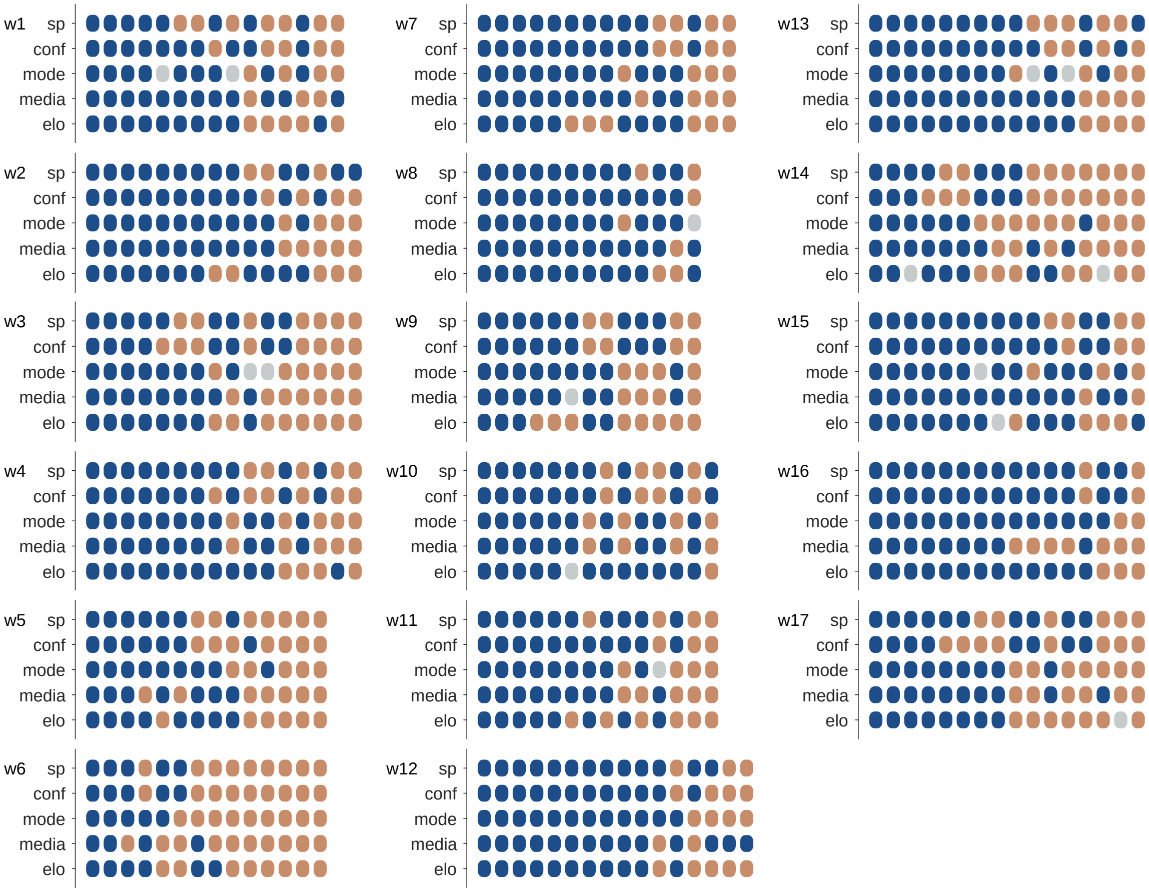

Figure 4. Correction for Lee et al. (2018, Figure 8). The accuracy of the predictions made by the surprising popular (sp), confidence-weighted tally (conf), mode of AMT participants (mode), mode of media experts (media), and Elo (elo) methods, for every game of the NFL season. Panels correspond to weeks of the seasons, rows to methods, and columns to games. Dark blue circles indicate a correct prediction; light orange circles indicate an incorrect prediction; and grey circles indicate neither team was favored.

> in our case, may be an important factor. Future work should try and isolate the type of expertise, or the types of games, that are likely to have the private knowledge or insight needed. As we mentioned, the ideal test is how well the surprisingly popular method performs when media experts provide meta-cognitive estimates of agreement, without being aware of the predictions others are making."

Acknowledgments

We are grateful to Marcellin Martinie for alerting us to the errors in the original analyses. The project page on the Open Science Framework project page at `https://osf.io/3kjmu/` has been updated to include corrected analyses for the surprisingly popular method applied to all 256 games.

References

Lee, M. D., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict NFL games. *Judgment and Decision Making*, *13*, 322–333.