Going with your gut: the (in)accuracy of forecast revisions in a football score prediction game

Going with your Gut: The (In)accuracy of Forecast Revisions in a Football Score Prediction Game*

Carl Singleton† J. James Reade‡ Alasdair Brown§

December 2019

Link to latest version

Abstract

This paper studies 150 individuals who each chose to forecast the outcome of 380 fixed events, namely all football matches during the 2017/18 season of the English Premier League. The focus is on whether revisions to these forecasts before the matches began improved the likelihood of predicting correct scorelines and results. Against what theory might expect, we show how these revisions tended towards significantly worse forecasting performance, suggesting that individuals should have stuck with their initial judgements, or their 'gut instincts'. This result is robust to both differences in the average forecasting ability of individuals and the predictability of matches. We find evidence this is because revisions to the forecast number of goals scored in football matches are generally excessive, especially when these forecasts were increased rather than decreased.

Keywords: Judgement revision; Prediction making; Sports forecasting
JEL codes: C53; C23; D84; Z2

*We are grateful for comments and advice from Alessandro Spiganti and the anonymous referees, as well as participants at the 39th Inernational Symposium on Forecasting in June 2019. This paper is based on data obtained from and analysed with the permission of Superbru, Sport Engage Ltd. Throughout the study, the anonymity of individual users of the Superbru prediction game was maintained. The use of these data does not imply the endorsement of the data owners in relation to the interpretation or analysis of the data.

†Corresponding author: c.a.singleton@reading.ac.uk; tel: +441183787864; Department of Economics, University of Reading, Whiteknights Campus, RG6 6UA, UK; with thanks to the Economic and Social Research Council (UK) for funding support under Grant No. ES/J500136/1.

‡University of Reading, j.j.reade@reading.ac.uk; Department of Economics, University of Reading, Whiteknights Campus, RG6 6UA, UK

§University of East Anglia, alasdair.brown@uea.ac.uk; School of Economics, University of East Anglia, UK

1 Introduction

Sports fans get pleasure from judging the outcomes of events in advance, in addition to possibly earning financial rewards if they choose to back these judgements in a betting exchange. For the most prominent and prestigious events, there is evidence of this in the ready supply of expert opinion in the media beforehand, which must be meeting a demand for insight on what is about to take place. Sport is therefore a particularly attractive context to study forecasting behaviour, or how individuals form expectations (Stekler et al., 2010). First, there is a ready supply of data on different types of forecasts for many similar events, for example: the probabilities implied by betting odds or prediction markets, and the picks (or point forecasts) from professional tipsters or everyday sports fans. Second, it is plausible that the outcomes of sports events are generally unaffected by the forecasts made of them, notwithstanding the possibility of corruption in prediction markets. This is less plausible in other contexts, such as company sales or financial securities forecasting. Third, the flow of information regarding an event and its potential outcomes is typically clean-cut, free and widespread in sport. For example, there is evidence that major news arriving within sports events are incorporated rapidly and completely among the participants of prediction markets (e.g. Croxson and Reade, 2014). Fourth, sports events have definitive outcomes and time horizons over which forecasts are being made, i.e. who wins or who loses and the end of the contest.

In this paper, we analyse the judgemental forecasts of 150 individuals who chose to play an online football score prediction game for all 380 matches in the 2017/18 season of the English Premier League (EPL), unquestionably the most watched domestic sports league in the world. The forecasts were of exact score outcomes, for example Manchester United to draw 1-1 playing at home against Liverpool. We also observe if, how and when the game players revised these forecasts before each match began. This allows us to ask an important question, the answer of which could apply to wider economic and social contexts: when forecasting the outcome of fixed events based primarily on judgement alone, do revisions or updates improve those forecasts? Or in other words, should individuals stick with their gut instincts? This is a question that has been addressed before in the laboratory (e.g. O'Connor et al., 1993, 2000; Lim and O'Connor, 1995). It has also been addressed in the field in the context of company sales forecasting (Lawrence and O'Connor, 2000). But, to the best of our knowledge, it has not been addressed in a context where large numbers of individuals voluntarily choose to make purely judgement-based forecasts and revisions (or not) at a time of their own choosing for many fixed events.For this reason, as well as those given above on the value of sports as a natural laboratory for studying forecasting behaviour, this paper provides significant new evidence on the effectiveness of judgement revisions. This evidence could have relevance to other contexts where judgemental forecasting explicitly takes place and which have real economic importance, such as in company management and planning (e.g. Edmundson et al., 1988), financial markets (e.g. De Bondt, 1993) and macroeconomic policy (e.g. Clements, 1995).

A reasonable working hypothesis would be that revising football scoreline forecasts as the match kick-off approaches should lead to improved forecast accuracy. There are numerous sources of information which become available to the forecaster over time and which are relevant to football match outcomes: not least betting odds (e.g. Forrest et al., 2005), other tipsters (e.g. Forrest and Simmons, 2000; Spann and Skiera, 2009) and the so-called 'wisdom of crowds' (e.g. Brown et al., 2018; Brown and Reade, 2018; Peeters, 2018). Furthermore, there are pre-match shocks which can significantly affect a team's abilities, like an injury to the star goalscorer.[1] However, there are plausible reasons why revisions to judgemental forecasts could lead to worse accuracy. First, there is evidence of bias among forecasts of football match results, such as the 'favourite-longshot' bias (e.g. Cain et al., 2000; Deschamps and Gergaud, 2007 and 'wishful-thinking' (Massey et al., 2011).[2] It is unknown and an outstanding empirical question, which we do not shed any direct light on here, whether these biases weigh more heavily on initial judgemental forecasts or subsequent revisions. Second, individuals may anchor on their previous forecast when making a revision, such that revisions are just white noise in the absence of relevant news (Clements, 1997). Third, the forecasting responses to salient new information can be excessive (Kahneman and Tversky, 1973; Tversky and Kahneman, 1974), overshooting and potentially even worsening the forecast, even if they are in the right direction towards the eventual outcome (De Bondt and Thaler, 1985, 1990; O'Connor et al., 1993; Lawrence and O'Connor, 2000).

We consider three main measures of forecast accuracy for football match scorelines: the percentage of correct scores, the percentage of correct results, and the points score achieved in the predictor game. Each game player revises at least one match forecast during the season and overall 6.3% of the 57,000 event forecasts are ever revised, with two-thirds of these only being done so once. Against all three measures, the overall forecast performance is worse in cases when the players ever revised their judgements compared with never. 9.2% of final scoreline forecasts actually happened and 9.3% were correct in cases which were never revised. This compares with 7.7% when a scoreline forecast had been revised at least once. In these cases that were revised, the initial forecasts would have gotten 9.2% of scores correct. Therefore, there is descriptive evidence suggesting that the game players would have been better off sticking with their first judgements rather than ever revising. We then use regression analysis and the panel structure of the data to confirm this result is robust, controlling for the time when forecasts and revisions were made relative to the events taking place, unobserved player forecasting ability and the predictability of each match. We find that scoreline forecasts that were revised had around 80% as great a likelihood of predicting a correct scoreline compared with those that weren't revised, and the players would

[1] Perhaps the most infamous example in recent sporting history occurred just 72 minutes before the 1998 World Cup final. The leading goalscorer in the tournament, Ronaldo, was inexplicably omitted from the starting lineup of Brazil. He was eventually reinstated and started the match, but was clearly not fully fit. Brazil went on to lose the match.

[2] Favourite-longshot bias is the tendency to observe in betting markets insufficient money backing the favourites to win, and too much backing the long-shots. Wishful-thinking bias refers to individuals backing the team to win with whom they have the greater affinity. Massey et al. (2011) looked at this in the context of a different brand of football: the US National Football League.

have done better had they stuck with their original judgements. When match results were revised as well as scorelines, players were also only 80% as likely to predict a correct result compared with cases where no revisions were made. Likewise, the prediction game points score was on average significantly lower in matches where players revised their initial forecasts. We also look at the goals scored forecast errors made by the players. There is some significant evidence that positive revisions (decreasing the number of goals scored by teams) are generally in the right direction and improve on these errors. However, negative revisions lead to greater forecast errors, though not significantly so at standard levels.

It is widely acknowledged that judgement plays a valuable role in many forecasting contexts, such as in company management and even macroeconomic forecasting (e.g. Fildes and Hastings, 1994; Fildes and Stekler, 2002)., and this paper contributes to the wider literature on judgemental forecasting (see for a summary Lawrence et al., 2006). In particular, it provides further evidence that new judgements may not improve on old ones when forecasters make revisions for fixed events. However, here the context is forecasting discrete and relatively low probability outcomes, as opposed to the continuous company sales forecasts or macroeconomic time series previously studied. It further contributes to the vast literature which generally investigates forecast efficiency, in particular when revisions are made (e.g. Nordhaus, 1987). This paper also adds to the growing list of studies which have used crowd-based data to analyse forecasting behaviour. Unlike the studies of Brown et al. (2018) and Peeters (2018), the members of the crowd we study are themselves directly making forecasts.

This paper also relates to the literature on using sports to study the practice of forecasting (see for a summary Stekler et al., 2010). One novelty here for football is that the game players, or tipsters, are forecasting exact scorelines. Several studies have developed statistical models to predict scorelines, though they have only tended to evaluate these models in terms of then forecasting match results (e.g. Dixon and Coles, 1997; Karlis and Ntzoufras, 2003; Goddard, 2005; Koopman and Lit, 2015; Boshnakov et al., 2017). However, no studies have yet looked at how successful individuals are at forecasting these relatively low probability and countably infinite variations on the outcomes of football matches.

Finally, the results in this paper contribute real-life evidence to the more general study of human decision making. When individuals update or revise beliefs, Tversky and Kahneman (1974) suggested potential sources of behavioural bias. One is that individuals may respond excessively to news which is salient. Another is that individuals are conservative in how they update their beliefs, anchoring on initial information and failing to adjust sufficiently to new information. For these reasons, researchers have traditionally emphasised the benefits of analytical decision-making to partially counter these biases (e.g Janis and Mann, 1977). We can only speculate about why the individuals that we study chose to revise their forecasts. One possibility is that they did so, typically after very small periods of time, and achieved worse outcomes, due to over-analysis, when they would have been better off sticking with their initial judgements.

The rest of the paper proceeds as follows: Section 2 describes the Superbru Premier League Predictor Game and the sample of player forecasts which we analyse; Section 3 measures the forecast accuracy in the game; Section 4 looks at how making revisions relates to subsequent forecast accuracy; Section 5 asks whether players' revisions are excessive; and Section 6 summarises.

2 Data

2.1 The Superbru Premier League Predictor Game

Superbru is an online platform for sports-based predictor and fantasy games. It is trademarked and managed by a South Africa based company, Sport Engage Ltd, which uses the platform for marketing purposes and generating revenue through advertising. Across all the games on the platform there are over 1.5 million registered users, who each play one or more of the games. These cover multiple sports, including cricket, rugby union, cycling, golf, motorsports and American football, among others. This paper focuses on the English Premier League Predictor Game. As of September 2018, there were 89,000 players of this game. Despite focusing on English football, the players are global, with just over a sixth based in the United Kingdom.[3]

The game is simple. In advance of each EPL fixture the player submits his forecast for the outcome of the match. There are 380 matches in a season between August and May of consecutive years. The forecasts are in the form of scorelines, for example 2-1, where the first number refers to the number of goals forecast to be scored by the team playing at home (their own stadium), and the second refers to the goals forecast for the opposition, who are playing away from home. There are 20 teams in the EPL, and each team plays all the others twice, home and away. All games end in regular time and there are three possible result outcomes: a home win, an away win or a draw. The aim of the game is to accumulate points through accurate forecasts of scorelines and results. Each forecast is awarded one of four possible amounts of points:

0 points: result is incorrect

1 point: result is correct

1.5 points: result is correct and forecast has a 'closeness' score ≤ 1.5 (see below)

3 points: scoreline is correct (and implies that result is correct too).

A forecast is determined to be close to the actual outcome of a match using Superbru's own metric of 'closeness', given by:

$$closeness_{ij} = |gd_j - \widetilde{gd}_{ij}| + \frac{|gs_j - \widetilde{gs}_{ij}|}{2}, \tag{1}$$

[3] Statistics obtained from www.superbru.com.

where i denotes a game player and j denotes a match. gd_j refers to the goal difference outcome of the match, measured as the home minus the away goals scored, and gs_j refers to the actual outcome sum of goals scored by both teams. The equivalent accented terms refer to the values implied by player i's forecast. In practice, a player will earn 1.5 closeness points if they forecast the correct result and the predicted scoreline has one goal more or less for one or both teams than the actual outcome of the match.

Players of the predictor game have several incentives to make accurate forecasts. First, there is a game leaderboard for each season. After each round of matches (typically 10), a player receives their global and nation rankings, both for that particular round and for the season as a whole up to that point. These are expressed in absolute terms and as percentiles. The player can earn eighteen different badges, corresponding to achievements in the game, such as for accuracy, streaks of correct results or scores and achieving a globally high rank at any point. Players can organise themselves into their own pools or leagues, perhaps competing against friends and family. Finally, there are prizes, including cash, for the number of correct scores out of 10 achieved in a round: 6, a Premier League jersey (cash value £50+); 7, £500; 8, £1,000; 9, £10,000; 10, £50,000. We asked Sport Engage Ltd for any evidence or views on why their users play the game and return to it week after week. They stated that their own marketing and customer survey data would suggest that the primary reason users keep coming back to the game is that they enjoy the social aspect of it, i.e. competing with their friends and family members directly in a well-designed setting. The company also said that their own research suggested users were to some extent motivated by the intrinsic utility of achieving correct forecasts, especially for the scoreline. However, we have no direct evidence to support these assertions and they should be treated with caution.[4] Generally, we would assume that the game users are aiming to maximise their scores, through correct scoreline picks. There is little incentive within the structure of the game to attempt a standout set of forecasts, by forecasting unlikely outcomes, such as high scoring matches. In expectation, this would not be rewarded by the cash prizes. Nor would it pay off in the overall leaderboard or pool rankings, as these are simply based on cumulative points throughout the season.

In practice, playing the game works as follows. Users access their game accounts and are immediately asked for forecasts of the upcoming round of EPL fixtures. Each match forecast is keyed in numerically and saved automatically after both the home and away goals are inputted. A user does not have to provide predictions on all the fixtures in a round. They can also make forecasts from the present day right up to the end of the season, for all future fixtures. No information is revealed to a player about what other players have forecast for a match unless they 'lock' their prediction. Locking a prediction prevents a player from further revising it, but they are then given information on the distribution of other players' predictions over possible scorelines for that match. If a prediction is not locked, then it can be revised at any point up to the start of

[4]The company also told us about the large amount of correspondence they receive from the users of this prediction game, who complain that correct score picks are not given sufficient weight in the scoring rule, and who enquire about or quibble their awarded closeness scores.

the match. After each round, players are sent an email update on their forecast accuracy, their game scores and relative rankings. Email reminders are also sent a couple of days before each round of matches, reminding players to make forecasts. The online game is available on PC and as an app for mobile devices. There are no monetary costs from playing the game or becoming a Superbru user. From our own experience, the process of making new forecasts for each round of fixtures is relatively costless, taking no more than a minute on a mobile phone, perhaps after being prompted by a reminder email. Revising those forecasts is similarly not time intensive. We presume that the game players are not using statistical modelling to make forecasts and are generally applying their best judgements, affected by whatever personal biases they might have and some of the information available to them, such as other forecasts from tipsters or betting odds, or news shocks, such as announcements on team starting line-ups.

2.2 A sample of game players

The administrators of the online prediction game provided us with all the forecasts made during the 2017/18 season by 150 players. These players were randomly sampled from the population of Superbru users who 'completed' the game, i.e who made forecasts for all 380 Premier League matches that season. We can distinguish the order in which players made multiple forecasts (revisions) for any particular match. To the nearest minute we observe the time at which each forecast was made. The 150 players are otherwise completely anonymous, and we have no information, for example, on team affiliations (which EPL club they are a fan of), ages or nationalities.

On the representativeness of the sample, first, we would speculate that the typical player of Superbru is a keener and more knowledgeable fan of EPL football than the typical person, or even the typical football fan, given they self-selected into playing. Second, users who 'completed' the game are likely to be particularly devoted to following the EPL, so we could speculate that they have better than normal knowledge and expertise about the events being forecast.[5] Third, to persist with the game, these players probably attain greater than normal utility from making forecasts and from getting them correct.

Table 1 shows some descriptive statistics for the sample of 57,000 final match forecasts in the sample, as well as for the actual outcomes of the 380 matches during the EPL 2017/18 season. For home goals scored forecasts, the players are more conservative than reality, though the average actual number of away goals scored reflects their forecasts closely. Overall, the mean total number of goals forecast to be scored by both teams in a match is 2.58, 0.1 goals lower than what occurred. The final forecast number of goals shows less variation across matches than in the outcomes. Looking at the frequency of scoreline forecasts, the game players tend to predict lower-scoring matches with one exception: only 1.5% of forecasts were for no goals in a game, compared with 8.4% of matches ending this way. This difference also accounts for why the game players

[5] We do not have good information on the distribution of the number of forecasts made during this season among EPL Predictor game players, other than from Superbru informing us that the drop-out rate from the game is high, and only a small fraction play throughout the season.

forecast fewer draws, 18.9%, compared with among outcomes, 26.1%. Taking account of this under-prediction of draws, there does not seem to be any clear bias in the forecasts for the team playing at home to win as opposed to the away team. The average time between the final scoreline forecast and the kick-off of a match is 81 hours, and the median is 55 hours. Some forecasts were made weeks or even months before the match took place.[6] In general, the game players are not waiting until right before kick-off to make their forecasts. This would frequently lead them to ignore what could be considered important information relating to match outcomes, such as which of a team's footballers are available for selection (typically announced 48-72 hours before a match) or even which footballers will take to the field (typically announced less than 1 hour before kick-off).

TABLE 1: Sample descriptive statistics

	Final predictions					Actual outcomes				
	Mean	Med.	Min.	Max.	St. dev.	Mean	Med.	Min.	Max.	St. dev.
Home goals	1.42	1	0	9	0.96	1.53	1	0	7	1.34
Away goals	1.15	1	0	6	0.97	1.15	1	0	6	1.18
Match goals	2.58	3	0	11	1.09	2.68	3	0	9	1.66
Hours to k-off	81.4	55.3	-0.0	3677	151.5					
Result (%):										
Home win			48.17					45.33		
Away win			32.94					28.42		
Draw			18.90					26.05		
Scores (%):										
0-0			1.5					8.4		
1-0			10.1					11.6		
2-0			10.7					7.1		
0-1			5.1					6.1		
0-2			6.0					3.9		
1-1			13.8					11.8		
2-1			15.1					8.4		
1-2			12.7					6.3		
other			25.0					36.3		
Players - n					150					
Matches - m					380					
Total - N					57,000					

[6] We do not have data on when players were sent email summaries of their forecasting performance or reminders to make forecasts. But from playing the game ourselves, reminders are sent around 2-3 days ahead of a round of fixtures (e.g. on a Wednesday or Thursday for matches on a Saturday to Monday), which is consistent with the timing patterns we observe.

2.3 Forecast revisions in the game

Game players can revise their forecasts as many times as they like and at any point before a match kicks off. We define these goals revisions by:

$$r_{ijk} = f'_{ijk} - f_{ijk} \ , \tag{2}$$

where $k \in \{h,a\}$ denotes whether the revision is made to the home or away goals scored. f_{ijk} denotes the new forecast following a revision and f'_{ijk} refers to some previous forecast. The vast majority of scoreline predictions by the players are never revised.

Before describing this relatively small number of revisions, first we consider the possibility that some observed revisions are simply mistakes (palpable errors or 'typos') made when players submit their forecasts. Appendix Table A1 shows the extent of all home and away goals revisions made in the sample between first and final scoreline predictions. There is suggestive evidence that some revisions are just players correcting their mistakes. For example, of the 3,781 forecasts ever revised, 31 revisions are by players decreasing the away goals scored forecast by 8 goals. As you can see from Table 1, the away team never actually scored 8 goals during the season, neither did any player predict they would score 8 with their final forecast. This compares with only single instances where players decreased their away goals forecasts by 7 and 9 goals. This would be consistent with 'fat finger' mistakes made when submitting forecasts on the Superbru platform from a mobile device using the ITU E 1.161 layout for a numbers keypad (i.e. with the 8 key above the 0). If in some cases players are correcting palpable forecast errors rather than making true revisions, and we are unable to distinguish these from one another, then we would anticipate estimating an improvement in forecast accuracy following an observed revision which is biased upwards. To partially address this, we arbitrarily label as 'mistakes' any scoreline revisions which involved the forecast number of home or away goals being revised by more than two goals, in either direction. This relabels 199 forecast cases as never having been revised and we drop these event forecasts from our following analysis unless stated otherwise. Table 2 and Appendix Tables A2-A3 describe the remaining 3,582 forecasts which were revised.

Table 2 presents statistics on the frequency of revisions over game players and matches. Every player makes at least one revision during the season. The average number of forecasts revised by players is 25, with the median number being 15 (4%). Conditional on revising the scoreline, players might also revise the result. A player did this on average 12 times during the season. We could expect that some matches are less likely to have forecasts revised than others. This could be due to the timing of a match kick-off or any especially pertinent new information, for example an injury during the warm-up to one team's goalkeeper or primary goal scorer. There is some variation across matches in the extent to which scorelines and results are revised, with a standard deviation of 3 forecasts for both. Deutscher et al. (2018) have demonstrated that newly promoted teams to a football league can be systematically underestimated on betting markets at the beginning of a season, which suggests that the Superbru game players could be less confident

when forecasting matches involving promoted teams. However, we find that the median number of forecast revisions among the 150 game players for matches involving at least one of the three promoted teams, 9, is the same as for matches not involving them, and the mean number of revisions is marginally lower in games involving promoted teams than those not involving them.

Typically, the time between the final and first forecast is around a minute. This could suggest that the majority of the revisions are not taking place following substantial changes in the game players' information sets. We can only speculate, but we might imagine that game players input their initial forecast, following which they look at the latest online betting odds on the match, or look for other information which might affect their judgement, such as news on team selection for the match. Alternatively, these revisions could simply be the result of changes to initial judgements without any new information. However, the mean time between the final and first forecast is almost two days, accounted for by a small number of players who revised their predictions after a long period of time. This suggests that in some cases players revise their forecasts following substantial changes in their information sets. The average home goals revision and the average away goals revision are a decrease of 0.15 and 0.13, respectively. As shown by Appendix Table A2, the clear majority (86%) of forecasts which are revised only involve the player revising either the home or away goals, with a slightly greater tendency to revise the away goals, and a tendency to make more conservative final forecasts. Appendix Table A3 also reflects this, showing if and how results were revised, with revisions tending more towards a win by the away side than the home. There is also evidence that players' first scoreline forecasts tend to favour draws more than after they are revised.

TABLE 2: Forecast revision descriptive statistics

	Mean	Med.	Min.	Max.	St. dev.
Number of forecasts revised per player:					
Score	23.86	15	1	215	28.73
Result	11.80	7	0	113	14.21
Number of forecasts revised per match:					
Score	9.43	9	2	22	3.37
Result	4.67	4	0	15	3.41
Hours between final and first forecast	45.97	0.02	0	1699	157.66
Home goals revision - r_{ijh}	-0.15	0	-2	2	0.84
Away goals revision - r_{ija}	-0.13	0	-2	2	0.91
Total number of forecasts revised:					
Score			3,582 (6.31%)		
Result			1,722 (3.12%)		

3 Forecast accuracy in the game

We consider four measures of forecast accuracy. Assuming that the players are playing the game to win and achieve a high ranking, we use the number of points awarded to each forecast in the

game, which we simply refer to as the 'Superbru' measure. Related, we measure the 'closeness' of forecasts, as defined by Equation (1), and where a lower amount implies a more accurate forecast. We also look at the percentage of correct scoreline and result forecasts, especially since the former determines the financial rewards available to game players.

The first row of Table 3 gives the forecast accuracy for all 57,000 final predictions in the sample of game players. The average Superbru score and closeness were 0.79 and 2.24, respectively. Over the whole season, the best player accumulated 350.5 points, whereas the worst player gained just 208. The median season score was 300.75, with a standard deviation of 19.5 points. Some matches were more 'predictable' than others.[7] There were matches during the season where not a single player forecast the correct scoreline or result. Similarly, there were matches where every player forecast the correct result. Just less than half of the 57,000 forecasts were in line with the actual match result. 9.16% of scorelines were correctly forecast by the players, with the best player achieving 13.42% correct picks, and the worst 5.52%.

TABLE 3: Final forecast accuracy

	Correct (%)		Superbru	
	Score	Result	Game score†	Closeness‡
Overall/Mean ($N = 57,000$)	9.16	49.55	0.79	2.24
Players ($n = 150$)*:*				
Median	9.21	49.87	300.75	2.23
Min.	5.52	35.53	208	2.00
Max.	13.42	55.00	350.5	2.66
St. dev.	1.56	2.65	19.52	0.12
Matches ($m = 380$)*:*				
Median	4.67	41.67	112.50	1.96
Min.	0.00	0.00	0.00	0.82
Max.	43.33	100.00	318.50	6.32
St. dev.	10.47	34.65	81.48	1.04

† Statistics for the the cumulative game score over matches for players and for the cumulative game score over players for matches. The Game score metric is described in Section 2.
‡ Statistics for the mean value of closeness for players throughout the season and for the mean value of closeness for matches over all players. Closeness for each forecast defined as per Equation (1).
Notes: *'Players'* displays distributional statistics over game players. *'Matches'* displays statistics over matches or events. These statistics include event forecasts where 'mistakes' were labelled.

3.1 Forecast revisions and accuracy

Table 4 summarises the forecast accuracy in the game depending on whether predictions were revised. For comparison, the first row repeats the overall accuracy for final forecasts, but now with the 'mistakes' dropped from the sample. The second row looks at the accuracy when either the match scoreline or result forecast were never revised by the player. The accuracy is marginally

[7] Superbru provide players with a measure of 'predictability' for matches in their performance summaries after each round.

higher than overall across all four measures considered. The third row of Table 4 describes the accuracy of final forecasts when they had at some point been previously revised by the game players. In these cases, 7.7% of correct scoreline forecasts were made, compared with 9.3% for forecasts which were never revised. Similarly, when results were revised, game players tended to have less forecasting success, and thus the Superbru game scores were also on average lower. However, the fourth row shows that players' first forecasts, i.e. before revision, would have on average achieved the correct score more often than their final forecasts: over 9% correct, significantly different from the final forecast accuracy in these cases, and at approximately the same rate as other forecasts which were never revised. Around a third of the match forecasts where scores were revised saw multiple changes. Therefore, we also look at the accuracy of the forecasts most recent to the final ones, still finding that over 9% of these scoreline picks would have been correct if players had stuck with their earlier judgements. In contrast, in the 1,722 cases were the result forecast was ever revised, corresponding to larger goals revisions, the previous forecasts of players were on average no more accurate than their final forecasts. The same can also be said for the average Superbru game score and closeness per match achieved by players with their initial and revised forecasts.[8]

TABLE 4: Revisions and forecast accuracy

	Correct (%)		Superbru (Mean)	
	Score	Result	Game score	Closeness
1. All final forecasts (56,801)	9.17	49.54	0.79	2.24
2. Final forecast, never revised (53,219/55,029)†	9.26	49.95	0.80	2.23
3. Final forecast, revised (3,582/1,772)†	7.71	36.63	0.74	2.28
4. First forecast (3,582/1,772)†	9.21	35.16	0.76	2.29
5. Previous to final forecast (3,582/1,772)†	9.02	35.21	0.76	2.32
Statistical significance of differences in means (p-values)‡:				
3. from 4.	0.02	0.47	0.35	0.71
3. from 5.	0.05	0.46	0.47	0.10

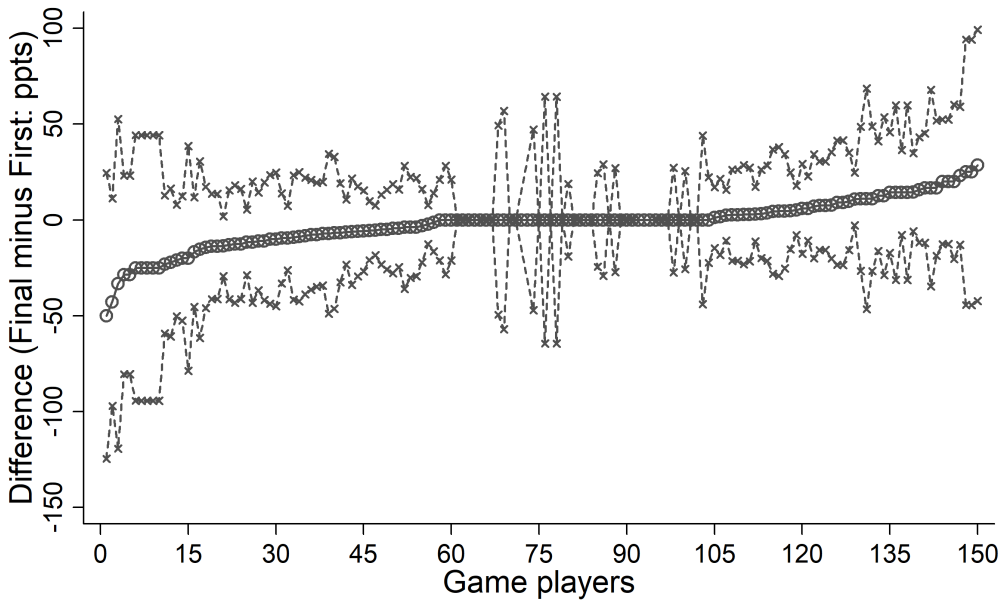
† The first and second numbers in parentheses give the number of score and result picks (never) revised, respectively.
‡ Paired two-sided t-test, where null is that the difference between mean accuracy is zero.
Notes: 'Mistakes' are excluded and not treated as final forecasts.

We also look at the heterogeneity over game players in whether or not forecast revisions tend to be associated with decreased accuracy. For each player, we consider the difference between the percentage of correct scorelines obtained with final and first forecasts, conditional on having made a revision. These differences for each player are plotted in Figure 1. There are some game players,

[8]The statistics in Table 4 excluded 'mistakes' from the set of revisions, i.e. scorelines where the home or away goals forecast were revised by more than 2 goals by a player. However, we find qualitatively similar results when we include these other 199 observations. For the percentage of correct scores, for example, we find the following, where the number in parentheses refers to the equivalent row in Table 4: (2), 9.26%; (3), 7.72%; (4), 9.21%; (5), 8.72%.

around a third, who have greater accuracy with their revised forecasts, though not significantly so. However, a greater number of game players on average experienced decreased forecast accuracy after making revisions.

FIGURE 1: Average difference between the accuracy of final and first forecasts when making revisions, by game player



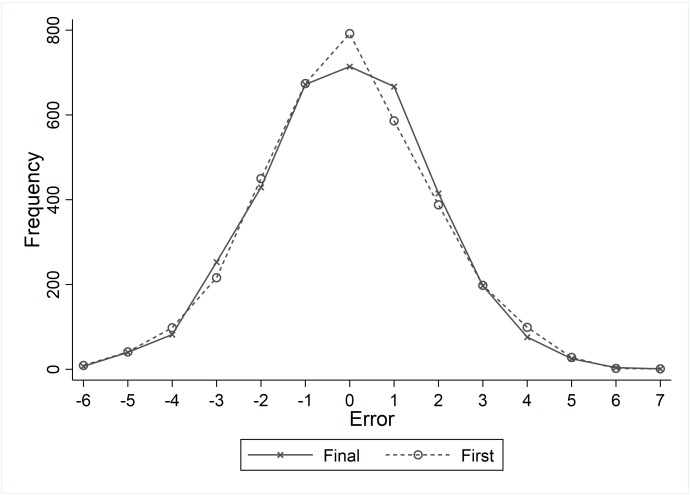
Notes.- author calculations using Superbru user data from 2017/18. Dashed lines and crosses show the estimated 95% confidence intervals for each game player.

Figure 2 plots the magnitude of errors made by players in cases where their forecasts were revised, comparing their final and first predictions. Figure 2A looks at the goal difference in matches. It shows that the number of cases when players achieved the correct goal difference was greater for their first forecasts. Similarly, Figure 2B shows that players generally increased the level of their forecast errors after making revisions, and that the first forecast was marginally more likely to predict the number of goals scored in a match correctly than the final forecast. Finally, Figure 2C plots the distribution of player closeness scores achieved as per Equation (1). This is in effect a weighted average of the forecast errors displayed in A and B, and a closeness score of zero corresponds to a correct scoreline forecast.

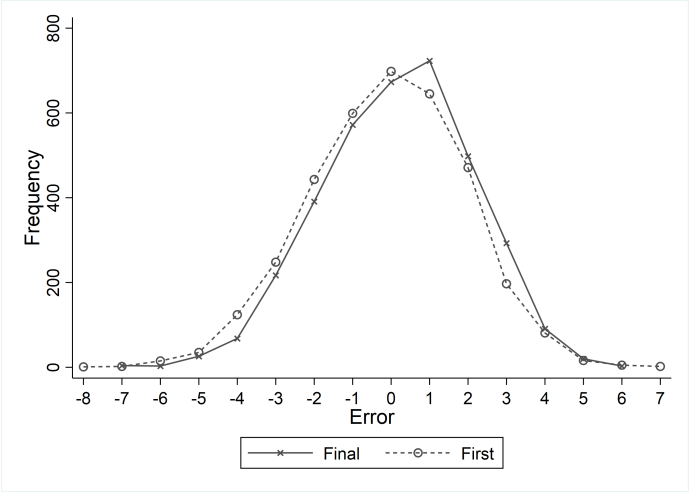
The statistics presented so far on forecast accuracy are certainly not causal evidence that when game players revise their judgements they perform worse. They are not even sufficient to descriptively suggest this is the case. First, it is plausible that the players who revise their forecasts more often are also those who tend to be weaker forecasters. Second, as described before, it is a fact that some matches are more difficult to predict than others. If players are more likely to revise their forecasts when matches are difficult to predict, then this could account for why revisions are associated with lower accuracy, as opposed to it mattering whether a forecast was revised at all. Third, there is substantial variation in this sample in terms of how long before the events the players made and revised their forecasts. These temporal differences could explain forecast

FIGURE 2: Distributions of final and first forecast errors: scoreline predictions which were ever revised only

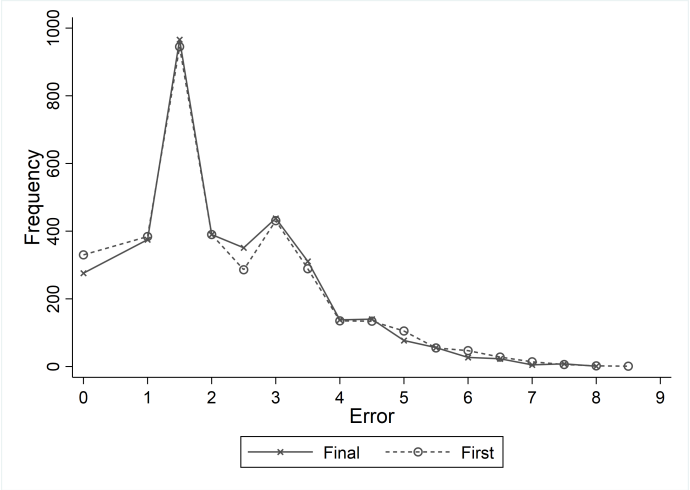
(A) Goal difference (Home minus Away goals)



(B) Total goals scored in a match



(C) Closeness - Equation (1)



Notes.- author calculations using Superbru user data from 2017/18. Goal difference and goals scored forecast errors are computed by subtracting the actual match outcome from the player's forecast. "Final" refers to the final forecasts ever revised by players. "First" refers to the first forecasts in those cases. Follows from Table A2 in so far as any score revisions by $> |2|$ away or home goals are excluded from the sample of forecast updates displayed here.

accuracy and be correlated with whether a pick is revised. We address these issues in the next section.

4 Should players revise their forecasts?

To address whether a game player revising their forecasts is associated with lower accuracy, we estimate a series of regression models, exploiting the panel nature of our sample. For correct score and result forecasts, since the outcome is binary, we estimate logistic regression models. In general, let the outcome of player's i's final forecast for game j be given by y_{ij}, where in the case of exact scores or results this takes a value of 1 if the forecast was correct and 0 otherwise. We look at logistic regression models of the following general form:

$$y^*_{ij} = revised_{ij}\lambda + \mathbf{x}'_{ij}\beta + \rho_{L(ij)} + \alpha_i + \mu_j + \varepsilon_{ij} \ , \qquad \varepsilon_{ij} \sim \text{Logistic}(0,1) \ , \tag{3}$$

where

$$y_{ij} = \begin{cases} 1; \text{ correct} & \text{if } y^*_{ij} > 0 \\ 0; \text{ incorrect} & \text{otherwise.} \end{cases}$$

In this equation, $revised_{ij}$ takes a value of 1 if the forecast was ever revised, with associated coefficient λ. The vector $\mathbf{x}_{ij}$ contains possible controls, which vary across players and matches, and β is a vector of coefficients. The set of covariates we can include in $\mathbf{x}_{ij}$ is limited. We consider information derived from the time when forecasts were made, to explore whether the time when predictions were made and revised explains their accuracy. We also derive variables which are proxies for how uncertain the outcome of a match was. For these we use the percentage of the 150 game users who made the modal scoreline or result forecast, reflecting the extent to which there was some 'consensus' about the most likely outcome of each match. $\rho_{L(ij)}$ gives a set of round fixed effects, where $l = L(ij) = 1,2,...,38$ and $L(ij)$ is a function indicating that player i made a forecast of match j taking place in round l of the EPL season. These round effects allow us to partially account for differences in the predictability of the matches being forecast, since each round consists of a different set of match-ups across teams. Individual player fixed effects are given by α_i. These control for the possibility that some players have higher forecasting ability than others. They also take care of other relevant omitted player-level fixed covariates, such as nationality, age, team affiliation, a player's average tendency to make forecasts a particular amount of time before matches kick-off, and the average extent to which they are a player who ever makes revisions. Similarly, match fixed effects are given by μ_j, which address the characteristics of matches which are common to all players making forecasts, such as the time and date of kick-off, which teams are taking part, the extent to which it was a difficult match to predict the outcome of, and whether players were on average more likely to revise their forecasts. Finally, ε_{ij} gives the error term, which is assumed to be logistically distributed. We cannot simultaneously estimate all the coefficients and effects in (3) as it is written above, since the round and match fixed effects would not be identified. Instead, we estimate variations of this regression model via maximum

likelihood with player-level or match-level cluster robust standard errors. In practice, we prefer a specification with round effects rather than match effects, alongside the player effects, since this reduces the amount of parameters that need to be estimated and thus exponentially reduces the time taken to estimate the model, as well as allowing us to retain a greater sample size as there is no outcome variable variation for some matches.

4.1 Results

4.1.1 Correct score

Table 5 shows our main regression estimates on whether players forecast correct scores with their final predictions. This is based on the full sample of 150 game players, since each makes at least one scoreline revision (Table 2). The results are presented as odds ratios. Column (I) shows the estimates of a standard logistic model based on Equation (3), where the player and match fixed effects are excluded and left in the error term. To address the possibility that when forecasts are made could be correlated with forecast accuracy and tendency to make revisions, covariates are included in the model for the number of days between the final forecast and kick-off, and this value squared.[9] However, these are generally not significant, suggesting that when a player makes their final scoreline forecast does not affect the likelihood of it turning out correct. Also included are round fixed effects, which are not displayed but are generally significant, and the number of days between the final forecast and the previous, i.e. the time taken to make a revision. This latter variable does have a small but insignificant effect on decreasing forecast accuracy through the longer players take to revise. The main results are given by the first row of Table 5, showing that players are only 83% as likely to forecast a correct scoreline after making revisions relative to when they make no revisions, with this effect being statistically significant at the 1% level. However, this is just a baseline result, consistent with the descriptive results of sample forecast accuracy described above in Section 3, since we have not yet addressed any unobserved player or match heterogeneity relevant to the accuracy of these forecasts. Without taking these into consideration, there could be plausible omitted variable bias affecting this result.

Column (II) of Table 5 shows our preferred specification, extending the model described by Column (I) by adding player fixed effects. It is well known that standard logistic regression estimates of models with fixed effects are inconsistent (Chamberlain, 1980). Even though the number of matches forecast within each group (by each player) is relatively large, 380, we nonetheless use conditional likelihood estimation, or what is now generally referred to as just fixed-effects logistic regression. The estimated negative and significant effect on scoreline accuracy of revising a forecast is then of a similar magnitude to before, with players 81% as likely to make a correct forecast if they had ever revised it compared with never. However, this result is much stronger than before as it is within player, i.e. it is robust to the possibility that worse forecasters revise more often. Column (III) adds a further explanatory variable to column (II),

[9] This is a quasi-continuous variable, since we know to the nearest minute when games kicked-off and forecasts were made.

TABLE 5: Logistic regression estimates for correct score, final forecasts: effects of having revised score (odds ratios)

	(I)	(II)	(III)	(IV)	(V)	(VI)
Revised	0.831***	0.811***	0.811***	0.819***		0.974
	(0.054)	(0.055)	(0.055)	(0.058)		(0.072)
(+1) Days to kick-off	0.998	1.010	1.006	0.996	1.009	
	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
(+1) Days to ... squared	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.072)	(0.000)	(0.000)
(+1) Days taken to revise	0.991	0.997	0.997	0.988	1.001	1.007
	(0.006)	(0.008)	(0.008)	(0.012)	(0.008)	(0.007)
(+1 ppt) Scoreline cons.			0.997			0.997
			(0.003)			(0.003)
Revised rounds 1-9					0.644***	
					(0.086)	
Revised rounds 10-19					0.782	
					(0.098)	
Revised rounds 20-29					0.905	
					(0.107)	
Revised rounds 30-38					0.906	
					(0.121)	
Round fixed effects	Yes	Yes	Yes	No	Yes	Yes
Player fixed effects	No	Yes	Yes	No	Yes	Yes
Match fixed effects	No	No	No	Yes	No	No
log pseudolikelihood	-17,097	-16,614	-16,614	-13,076	-16,612	-16,756
N	56,801	56,801	56,801	44,836	56,801	56,801

***,** indicate significance from one at 1% and 5% levels, respectively, two-sided tests. Cluster-robust standard errors estimated and displayed in parentheses, with (150) clusters at the player level, except for (IV), where (301) clusters are at the match level.
(I)-(V): dependent variable is whether a player's final forecast was correct.
(VI): dependent variable is whether a player's first forecast was correct.

serving as a measure of the extent to which there was some consensus among users about the match scoreline outcome. This variable is insignificant, and the results are unchanged. In Column (IV), we drop the player and round effects and estimate the model with match fixed effects. There were 79 matches during the season with no outcome variation because no player forecast the correct score, and so 11,850 observations are excluded from the estimation. In this case, we also find that revising a forecast is associated with around an 80% likelihood of predicting a correct score relative to revising it. This suggests that the lower predictability of some matches, which may tend towards more revisions being made, is not driving the fact that in these data forecasters perform worse when they do revise than when they don't. Finally, in Column (V), we return to the preferred specification as per Column (II), but allow for the marginal effects of revisions on the final forecast accuracy to vary throughout the season. The results show that revisions are associated with decreased scoreline forecast accuracy throughout the season, but only significantly so at the 5% level during the first quarter.

However, none of the above is evidence that the players should not have revised their forecasts. We construct a counterfactual outcome variable to investigate this. Instead of y_{ij} referring to the forecasting outcome of a player's final prediction, we instead look at the forecasting outcome of their first prediction. Column (VI) estimates the equivalent fixed effects logistic model as in Column (III), except that the dependent variable is whether or not the first scoreline forecast made by a player turned out to be correct. The estimated odds ratio relating to whether that forecast was subsequently revised is insignificantly different from 1, implying that the scoreline accuracy of a player's first pick is unrelated to whether it was revised. In other words, players were just as good forecasters with their first judgements independent of whether they subsequently made any revisions. This compares with the finding that player final forecast performance was significantly worse if revisions were made. Taken together, this suggests that players would have performed better if they had stuck with their initial judgements or gut instincts about match scorelines. For robustness, Appendix Table A4 presents comparable results to Table 5, where the additional 199 cases in which home or away goals scored forecasts were revised by more than 2 goals are re-classified as revisions rather than 'mistakes'. The results are approximately unaffected by including these cases.

4.1.2 Correct result

We also estimate the equivalent regression models for correct result forecast outcomes as for correct scores above. As well as the dependent variable being different, these models differ in that $revised_{ij}$ now concerns whether the result was ever revised by the player, not just the scoreline. Table 6 presents these model estimates for 146 of the game players; 4 players made no results revisions throughout the season.[10] We also find for results that revising forecasts is associated with significantly worse accuracy. Without controlling for player or match fixed effects, revisions lead to final forecasts being 59% as likely to be correct (Column I). Conditional on individual players' forecasting abilities, this figure falls marginally to 58% (Column II). In the latter set of results, there is no significant evidence that the time taken to kick off or the time taken to make a revision affect the likelihood of a correct result prediction. However, once we address the possibility that for some matches the result is less predictable than for others, by adding a variable capturing the extent to which there was a consensus about the result outcome of the match, we find that revising forecasts is associated with 83% as much of a chance of getting a correct result than without ever having revised them (Column III). This demonstrates how relevant heterogeneity in the events being forecast could have led us to overestimate the negative effects of making result revisions through omitted variable bias. In fact, a one percentage point increase in the number of users who forecast the most common outcome is associated with a 2% increased likelihood within user of a correct result forecast. Further, an increase in the time taken to make a revision significantly affects the likelihood of a correct prediction in this specification. Column (IV) further emphasises the need to account for the predictability of the match, by controlling for match fixed

[10] These 4 players did not have a significantly different rate of success in forecasting the result than the other 146 players (p=0.88, two-tailed t-test).

effects. Column (V) returns to the player fixed effects model and allows the effects of revisions on forecast accuracy to vary with each quarter of the season. A revision is significantly associated with a decreased likelihood of a correct result forecast in the first half of the season at the 1% level. We also look at whether the accuracy of a player's first result forecast is related to whether they then subsequently made a revision. Unlike for scoreline forecasts, we find this is significantly the case (Column VI). However, the time taken to make a revision is associated with a significantly increased likelihood that the original result forecast was correct. When it comes to the correct result, game players are just as well off revising their result forecasts than if they don't, so long as they do so quickly. However, as they wait longer, there is evidence that they would have been better off sticking with their original forecast.

TABLE 6: Logistic regression estimates for correct result, final forecasts: effects of having revised result (odds ratios)

	(I)	(II)	(III)	(IV)	(V)	(VI)
Revised	0.592***	0.578***	0.790***	0.828**		0.691***
	(0.033)	(0.035)	(0.047)	(0.067)		(0.046)
(+1) Days to kick-off	1.000	1.006	0.997	0.996	0.997	1.006
	(0.002)	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)
(+1) Days to ... squared	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
(+1) Days taken to revise	0.771***	0.862	0.684***	0.410***	0.712***	1.372***
	(0.069)	(0.080)	(0.074)	(0.137)	(0.083)	(0.156)
(+1 ppt) Result cons.			1.028***		1.028***	1.028***
			(0.000)		(0.000)	(0.000)
Revised rounds 1-9					0.751***	
					(0.081)	
Revised rounds 10-19					0.622***	
					(0.069)	
Revised rounds 20-29					1.003	
					(0.116)	
Revised rounds 30-38					0.842	
					(0.109)	
Round fixed effects	Yes	Yes	Yes	No	Yes	Yes
User fixed effects	No	Yes	Yes	No	Yes	Yes
Match fixed effects	No	No	No	Yes	No	No
log pseudolikelihood	-36,945	-36,398	-34,292	-21,356	-34,287	-36,391
N	55,282	55,282	55,282	50,384	55,282	55,282

***,** indicate significance from one at 1% and 5% levels, respectively, two-sided tests. Cluster-robust standard errors estimated and displayed in parentheses, with (146) clusters at the user level, except for (IV), where (337) clusters are at the match level.
(I)-(V): dependent variable is whether a player's final forecast was correct.
(VI): dependent variable is whether a player's first forecast was correct.

4.1.3 Superbru game score

Finally, we estimate a set of linear regression models, comparable to the logistic ones described above, where the dependent variable is the number of points a final forecast is awarded according

to the Superbru EPL Predictor Game. To some extent, this provides a more general measure of forecast accuracy. In this case, the variable of interest, $revised_{ij}$, measures whether a scoreline forecast was revised. Table 7 presents the results. Not controlling for player or match heterogeneity, revising a forecast is associated with achieving a 0.04 points significantly lower game score per match, (Column I). Conditional on the individual players themselves, this effect remains significant and the magnitude of the effect increases to 0.05 points (Column II). Given that on average the players in this sample revised 24 of their 380 forecasts during the season, the overall effect of revisions on the season-long forecast performance of players is nonetheless small. Within player, there is also evidence that the forecast score increases as players predict at greater horizons from the match kick-off, though decreasingly so with the length of time. Also, the time taken to make a revision tends to decrease the forecast score. The estimated negative effect of making a revision on the match game score is 0.04 points when we additionally control for the predictability of matches, using the same measure of consensus about the match result outcome as used above for correct result forecasts (Column III). The negative effect in cases where forecasts were revised remains significant also within matches (Column IV). The marginal effect of revision on the Superbru score for players is negative throughout the season, though only significantly so at the 5% level in the second quarter (Column V). Like the accuracy of scoreline forecasts, the counterfactual game score achieved by players' first forecasts, before any possible revisions, is not negatively related to whether that forecast was subsequently revised (Column VI). In other words, on average the players would not have performed significantly worse on the game had they stuck with their initial judgements.

Taken together, we conclude from the results in this section that revisions negatively affected a player's performance on the Superbru game leaderboards and their likelihood of winning cash prizes, as their predictions prior to revision were generally better, especially with respect to getting correct the exact scoreline.

5 Are goals forecast revisions excessive?

To understand our findings from the previous section, we investigate how the magnitude of the goals revisions made by game players relates to their forecast errors. One potential explanation for the fact that revisions in this context do not appear to improve accuracy is that they could in some sense be excessive. This has been found to be the case in laboratory experiments (e.g. O'Connor et al., 1993; Lim and O'Connor, 1995) and in the field among company sales forecasters (Lawrence and O'Connor, 2000).

To test whether this is the case here, we estimate variations on the following model using least squares:

$$e_{ijk} = r_{ijk}\gamma + r_{ij-k}\psi + \mathbf{x}'_{ij}\beta + \rho_{L(ij)} + \alpha_i + \mu_j + \varepsilon_{ijk} \ , \tag{4}$$

where the dependent variable e_{ijk} is the forecast error made by player i in match j for home or away goals, denoted by k. These errors are computed by subtracting a player's forecast number

TABLE 7: Linear regression estimates for Superbru match score: effects of having revised score

	(I)	(II)	(III)	(IV)	(V)	(VI)
Revised	-0.043***	-0.049***	-0.039**	-0.035**		-0.033
	(0.015)	(0.017)	(0.016)	(0.014)		(0.019)
(+1) Days to kick-off	-0.0001	0.003***	0.0003	-0.001	0.0003	0.0007
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
(+1) Days to ... squared	0.000003	-0.00002***	-0.00000	0.000001	-0.00000	-0.00000
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
(+1) Days taken to revise	-0.005***	-0.003**	-0.005***	-0.006***	-0.004***	0.001
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
(+1 ppt) Result cons.			0.007***		0.007***	0.007***
			(0.000)		(0.000)	(0.000)
Revised rounds 1-9					-0.056	
					(0.029)	
Revised rounds 10-19					-0.061**	
					(0.029)	
Revised rounds 20-29					-0.024	
					(0.031)	
Revised rounds 30-38					-0.012	
					(0.035)	
Round fixed effects	Yes	Yes	Yes	No	Yes	Yes
Player fixed effects	No	Yes	Yes	No	Yes	Yes
Match fixed effects	No	No	No	Yes	No	No
R^2	0.031	0.031	0.058	0.334	0.056	0.031
N	56,801	56,801	56,801	56,801	56,801	56,801

***,** indicate significance from one at 1% and 5% levels, respectively, two-sided tests. Cluster-robust standard errors estimated and displayed in parentheses, with (150) clusters at the player level, except for (III), where (380) clusters are at the match level.
(I)-(IV), (VI):: dependent variable is a player's match Superbru score from their final forecast.
(V): dependent variable is a player's match Superbru score from their first forecast.

of goals scored in a match from the actual number scored in the match outcome. The number of goals revised by is r_{ijk}, as defined by Equation (2). We might anticipate some correlation of the revision made to home (away) goals and the away (home) goals forecast error and revisions, and so we also include this in the model, given by r_{ij-k}. As before in the logistic regression models described by Equation (3), we also allow for other covariates in $\mathbf{x}_{ij}$, round effects and player or match fixed effects. Conditional on the other variables in the model besides r_{ijk}, the coefficient of interest is γ.

Note that we can re-write (4) as:

$$e_{ijk} = g e'_{ijk} + z_{ijk} \ ,$$

where e'_{ijk} is the error implied by the forecast before revision, $g = \frac{-\gamma}{1-\gamma}$, and $z_{ijk} = \frac{1}{1-\gamma}(r_{ij-k}\psi + \mathbf{x}'_{ij}\beta + \rho_{L(ij)} + \alpha_i + \mu_j + \varepsilon_{ijk})$ gives some part of the forecast errors accounted for by other factors or randomness. Therefore, if $\gamma \geq 1/2$, then $|g| \geq 1$ and the revised forecasts are on average worse, conditional on z_{ijk}, which does not necessarily have mean zero. If $\gamma < 0$, then $0 < g < 1$ and revisions are generally in the right direction, improving the forecasts conditional on z_{ijk}. However, if $0 < \gamma < 1/2$, then $-1 < g < 0$ and the revisions are on average in the right direction and improve

forecasts, but are excessive, overshooting and reversing the sign of the eventual forecast errors conditional on z_{ijk}.

We also look at whether the effects on final forecast errors are symmetric, in practice estimating variations on the following model:

$$e_{ijk} = \{\mathbf{1}_{r_{ijk}>0}\} r_{ijk}\gamma_p + \{\mathbf{1}_{r_{ijk}<0}\} r_{ijk}\gamma_n + r_{ij-k}\psi + \mathbf{x}'_{ij}\beta + \rho_{L(ij)} + \alpha_i + \mu_j + \varepsilon_{ijk} \ , \qquad (5)$$

where γ_p and γ_n now measure the effects of positive and negative revisions, respectively.

5.1 Results

Since the majority of forecasts which are revised are only done so once, we initially regress final forecast errors on the goals revisions since the first forecast of a match. We do so for the 2,076 and 2,290 cases where players made at least a one goal home or away scoreline revision, respectively, i.e. there are no zero values included for the goals revisions in r_{ijk}. Column (I) of Table 8 reports estimates of Equation (4) separately for home and away goals errors, with player and round fixed effects, but initially omitting the other goals revisions, r_{ij-k}, and the match fixed effects. The first row reports the estimates of γ in each case. The table reports the results of testing whether the coefficient estimate is significantly different from 0 or 1/2. The results do suggest that players' goals revisions are excessive, with values of γ being 0.48 and 0.40 for home and away goals revisions, respectively. Both values are clearly significantly different from zero, but only the away goals coefficient estimate is significantly different from 1/2 at standard levels. From these estimates, we find that goals revisions do in fact reduce the magnitude of errors in away goals forecasts, but not significantly so for home goals. This is not necessarily inconsistent with our previous finding for correct scores in Section 4, that players would be better off sticking with their first forecasts rather than revising them, since this previous finding was based on a binary outcome and not concerned with the magnitude of the errors being made.

Column (II) of Table 8 reports further results from estimating Equation (5), allowing the effects for positive (revising down) and negative (revising up) goals revisions to differ. We find that γ_p is positive but not significantly different from zero at standard levels of significance, indicating that when players revise down the goals scored by teams this generally improves their forecasts and these revisions are not excessive. We find that γ_n is greater than 1/2, though not significantly so. However, the coefficients for both home and away goals revisions are significantly different from zero, indicating that negative goals revisions tend to be excessive. Column (III) adds r_{ij-k} to the model but these other goals revisions are insignificant. Column (IV) estimates the model with match fixed effects instead of player and round effects. Qualitatively the result is unchanged; on average home and away positive goals revisions are in the correct direction, tending to improve the forecast, whereas negative revisions lead to greater forecast errors, though not significantly so.

TABLE 8: Linear regression estimates for home or away goals scored forecast errors from final prediction: effects of goals revisions

	(I)		(II)		(III)		(IV)	
	Home	Away	Home	Away	Home	Away	Home	Away
Revisions ($\widehat{\gamma}$)	0.48***	0.40***†††						
	(0.03)	(0.02)						
Revisions, +ve ($\widehat{\gamma}_p$)			0.20	0.15†††	0.20	0.15†††	0.38***	0.14***†††
			(0.15)	(0.10)	(0.16)	(0.10)	(0.06)	(0.05)
Revisions, -ve ($\widehat{\gamma}_n$)			0.74***	0.63***	0.74***	0.64***	0.55***	0.68***
			(0.14)	(0.08)	(0.14)	(0.08)	(0.06)	(0.05)
Days to kick-off	0.0168	-0.0229	0.0156	-0.0226	0.0156	-0.0228	0.0066	-0.122
	(0.015)	(0.012)	(0.015)	(0.012)	(0.015)	(0.012)	(0.007)	(0.007)
Days to ... squared	-0.0001	0.0004**	-0.0000	0.0004**	-0.0000	0.0004**	-0.0002**	0.0002
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.0001)
Days taken to revise	-0.0228	-0.0162	-0.0209	-0.0136	-0.0209	-0.0127	-0.0307***	-0.0060
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)		(0.006)	(0.005)
Days taken ... squared	0.0005***	0.0002	0.0004***	0.0002	0.0005***	0.0001	0.0005***	0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Revision, other ($\widehat{\psi}$)					-0.002	0.042	0.025	0.04
					(0.044)	(0.033)	(0.020)	(0.022)
Round fixed effects	Yes		Yes		Yes		No	
Player fixed effects	Yes		Yes		Yes		No	
Match fixed effects	No		No		No		Yes	
R^2	0.205	0.194	0.206	0.199	0.206	0.199	0.869	0.850
N	2,076	2,290	2,076	2,290	2,076	2,290	2,076	2,290

***,** indicate significance from zero at 1%, and 5% levels, respectively, two-sided tests. Cluster-robust standard errors estimated and displayed in parentheses, with (140/148) clusters at the player level, except for (III), where (379) clusters are at the match level.

For "Revision", †††,†† indicate significance from one half at 1%, and 5%, respectively, two-sided tests.

(I)-(IV): dependent variable is goals final forecast error (actual minus forecast)

Some players make multiple revisions. But the results on excessive goals revisions described above were based only on payers' first forecasts relative to their last for any given event. Therefore, we also estimate the exact same regression models as before, but now let r_{ijk} and r_{ij-k} refer to the goals revisions between a player's final scoreline forecast and their previous one for each match. The results are shown in Appendix Table A5. Qualitatively, the effects of home and away goals forecast revisions made between the final and previous forecast are the same as for revisions made between the final and first forecasts. On average, positive goals revisions reduce the magnitude of errors but are still generally excessive. However, the overshooting of the actual outcome scoreline is less than for the first forecast, for both home and away goals, conditional on the player or the match concerned. When players increase their goals scored forecasts, making less conservative judgements, the magnitude of forecast errors are generally increased.

6 Summary and further discussion

In this paper, we have analysed the forecasting performance of individuals who each applied their judgement to predict the outcomes of many fixed events. The context of this analysis was the scoreline outcomes of professional football matches. We found that when individuals made revisions their likelihood of predicting a correct scoreline, which they achieved around 9% of the time when never making a revision, significantly decreased. The same applied for forecast revisions to the result outcomes of matches. Not only were these findings robust to unobserved individual forecasting ability and the predictability of events, but also there is evidence that performance would have improved had initial judgements been followed.

As already mentioned, these results have some similarities with those found previously in the behavioural forecasting literature. One explanation could be that game players anchor their beliefs, expectations and, consequently, their forecasts on past or initial values. However, this behaviour would not be consistent with our finding that on average forecasters made revisions which not only improved on their goals scored forecast errors but which were also excessive.

There are several areas for further research, which could be explored with extensions of the dataset used here. First, it appears to be a relatively open question as to how sources of bias among sports forecasters interact with how they make revisions, such as the well-known favourite-longshot bias. Second, players of the forecasting game studied here do reveal which EPL team they have the greatest affinity for, though we are yet to observe this information ourselves. It is an interesting question as to whether any wishful-thinking by the players manifests itself more greatly before or after they revise their forecasts. Third, an aspect which could be studied from these current data is whether players improve their forecasts over time, and if they learn how to play more to the rules of the game itself, which should lead them to favour more conservative goals forecasts. Fourth, these results concern a selective random sample of players who "completed" the game. These are likely to be individuals who extract significant utility from making forecasts of football match scorelines, who are thus more likely to return to their initial forecasts and make

revisions. It would be interesting whether more casual forecasters are better at sticking with their gut instincts or better off from doing so. Finally, our results suggested an innovation to the game which could improve the crowd's forecasting accuracy and which could be easily tested: before making forecasts, some of the game players could be informed that sticking with their initial judgement, or gut instinct, is likely to improve their chances of picking a correct score.

References

Boshnakov, G., T. Kharrat, and I. McHale. 2017. "A bivariate Weibull count model for forecasting association football scores." *International Journal of Forecasting*, 33(2): 458–466.

Brown, A., D. Rambaccussing, J. J. Reade, and G. Rossi. 2018. "Forecasting with Social Media: Evidence from Tweets on Soccer Matches." *Economic Inquiry*, 56(3): 1748–1763.

Brown, A., and J. J. Reade. 2018. "The wisdom of amateur crowds: Evidence from an online community of sports tipsters." *European Journal of Operational Research*, forthcoming.

Cain, M., D. Law, and D. Peel. 2000. "The favourite-longshot bias and market efficiency in uk football betting." *Scottish Journal of Political Economy*, 47(1): 25–36.

Chamberlain, G. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies*, 47(1): 225–238.

Clements, M. P. 1995. "Rationality and the Role of Judgement in Macroeconomic Forecasting." *The Economic Journal*, 105(429): 410–420.

Clements, M. P. 1997. "Evaluating the Rationality of Fixed-event Forecasts." *Journal of Forecasting*, 16(4): 225–239.

Croxson, K., and J. J. Reade. 2014. "Information and Efficiency: Goal Arrival in Soccer Betting." *The Economic Journal*, 124(575): 62–91.

De Bondt, W. F. M., and R. Thaler. 1985. "Does the Stock Market Overreact?." *Journal of Finance*, 40(3): 793–805.

De Bondt, W. F. M., and R. Thaler. 1990. "Do Security Analysts Overreact?." *American Economic Review*, 80(2): 52–57.

De Bondt, W. P. 1993. "Betting on trends: Intuitive forecasts of financial risk and return." *International Journal of Forecasting*, 9(3): 355 – 371.

Deschamps, B., and O. Gergaud. 2007. "Efficiency in Betting Markets: Evidence from English Football." *Journal of Prediction Markets*, 1(1): 61–73.

Deutscher, C., B. Frick, and M. Ötting. 2018. "Betting market inefficiencies are short-lived in German professional football." *Applied Economics*, 50(30): 3240–3246.

Dixon, M. J., and S. C. Coles. 1997. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market." *Applied Statistics*, 47(3): 265–280.

Edmundson, B., M. Lawrence, and M. O'Connor. 1988. "The use of non-time series information in sales forecasting: A case study." *Journal of Forecasting*, 7(3): 201–211.

Fildes, R., and R. Hastings. 1994. "The Organization and Improvement of Market Forecasting." *Journal of the Operational Research Society*, 45(1): 1–16.

Fildes, R., and H. Stekler. 2002. "The state of macroeconomic forecasting." *Journal of Macroeconomics*, 24(4): 435 – 468.

Forrest, D., J. Goddard, and R. Simmons. 2005. "Odds-setters as forecasters: The case of English football." *International Journal of Forecasting*, 21(3): 551–564.

Forrest, D., and R. Simmons. 2000. "Forecasting sport: the behaviour and performance of football tipsters." *International Journal of Forecasting*, 16(3): 317–331.

Goddard, J. 2005. "Regression Models for Forecasting Goals and Match Results in Association Football." *International Journal of Forecasting*, 21 331–340.

Janis, I. L., and L. Mann. 1977. *Decision making: A psychological analysis of conflict, choice, and commitment.*. Free press.

Kahneman, D., and A. Tversky. 1973. "On the psychology of prediction." *Psychological Review*, 80(4): 237–251.

Karlis, D., and I. Ntzoufras. 2003. "Analysis of sports data by using bivariate Poisson models." *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3): 381–393.

Koopman, S. J., and R. Lit. 2015. "A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League." *Journal of the Royal Statistical Society Series A*, 178(1): 167–186.

Lawrence, M., P. Goodwin, O. Marcus, and D. Onkal. 2006. "Judgmental forecasting: A review of progress over the last 25 years." *International Journal of Forecasting*, 22(3): 493–518.

Lawrence, M., and M. O'Connor. 2000. "Sales forecasting updates: how good are they in practice?." *International Journal of Forecasting*, 16(3): 369 – 382.

Lim, J. S., and M. O'Connor. 1995. "Judgemental adjustment of initial forecasts: Its effectiveness and biases." *Journal of Behavioral Decision Making*, 8(3): 149–168.

Massey, C., J. P. Simmons, and D. A. Armor. 2011. "Hope over experience: Desirability and the persistence of optimism." *Psychological Science*, 22(2): 274–281, PMID: 21228135.

Nordhaus, W. D. 1987. "Forecasting Efficiency: Concepts and Applications." *The Review of Economics and Statistics*, 69(4): 667–674.

O'Connor, M., W. Remus, and K. Griggs. 1993. "Judgemental forecasting in times of change." *International Journal of Forecasting*, 9(2): 163–172.

O'Connor, M., W. Remus, and K. Griggs. 2000. "Does updating judgmental forecasts improve forecast accuracy?." *International Journal of Forecasting*, 16(1): 101–109.

Peeters, T. 2018. "Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer results." *International Journal of Forecasting*, 34(1): 17–29.

Spann, M., and B. Skiera. 2009. "Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters." *Journal of Forecasting*, 28(1): 55–72.

Stekler, H., D. Sendor, and R. Verlander. 2010. "Issues in sports forecasting." *International Journal of Forecasting*, 26(3): 606–621.

Tversky, A., and D. Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–1131.

Appendix A. Additional tables

TABLE A1: Goals revisions, "mistakes" included: first forecast minus final

Home goals revision - r_{ijh}	Away goals revision - r_{ija}														
	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	Total
4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
3	0	0	2	1	7	0	1	1	0	0	0	0	0	0	12
2	0	0	9	22	70	16	3	2	0	0	0	0	1	0	123
1	0	4	19	229	748	143	15	2	1	0	0	0	0	0	1,161
0	2	11	108	797	N/A	542	59	67	6	4	3	1	29	1	1,630
1	0	3	23	120	446	146	7	3	0	0	0	0	1	0	749
2	0	0	9	10	28	8	5	1	0	0	0	0	0	0	61
3	1	1	0	0	18	0	2	1	0	0	0	0	0	0	23
4	0	0	0	2	3	0	0	0	1	0	0	0	0	0	6
6	0	0	0	1	2	0	0	0	0	0	0	0	0	0	3
7	0	0	0	0	2	1	0	0	0	0	0	0	0	0	3
8	0	0	1	3	4	0	0	0	0	0	0	0	0	0	8
9	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
Total	3	19	171	1,185	1,329	857	92	77	8	4	3	1	31	1	3,781

TABLE A2: Goals revisions, $> |2|$ excluded: first forecast minus final

Home - r_{ijh}	Away - r_{ija}					
	-2	-1	0	1	2	Total
-2	9	22	70	16	3	120
-1	19	229	748	143	15	1,154
0	108	797	N/A	542	59	1,651
1	23	120	446	146	7	742
2	9	10	28	8	5	60
Total	168	1,178	1,437	855	89	3,582

Notes: See Appendix Table A1 for all user revisions.

TABLE A3: Updated result forecasts following goals revisions

Original	Updated			
	Away	Draw	Home	Total
Away	670	306	121	1,097
Draw	370	127	451	948
Home	177	347	1,013	1,537
Total	1,217	780	1,585	3,582

Notes: Follows from Table A2 in so far as any score revisions by $> |2|$ away or home goals are excluded from the sample of forecast updates.

TABLE A4: Logistic regression estimates for correct score, final forecasts: effects of having revised score (odds ratios) — all revisions including 'mistakes'

	(I)	(II)	(III)	(IV)
Revised	0.822***	0.803***	0.808***	0.975
	(0.051)	(0.053)	(0.054)	(0.072)
(+1) Days to kick-off	0.998	1.006	0.996	1.009
	(0.004)	(0.005)	(0.005)	(0.005)
(+1) Days to ... squared	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)
(+1) Days taken to revise	0.998	1.004	0.996	1.008
	(0.004)	(0.008)	(0.012)	(0.007)
Round fixed effects	Yes	Yes	No	Yes
Player fixed effects	No	Yes	No	Yes
Match fixed effects	No	No	Yes	No
log pseudolikelihood	-17,151	-16,669	-13,134	-16,756
N	57,000	57,000	45,150	57,000

***,** indicate significance from one at 1% and 5% levels, respectively, two-sided tests. Cluster-robust standard errors estimated and displayed in parentheses, with (150) clusters at the player level, except for (III), where (301) clusters are at the match level.
(I)-(III): dependent variable is whether a player's final forecast was correct.
(IV): dependent variable is whether a player's first forecast was correct.

TABLE A5: Linear regression estimates for home or away goals scored forecast errors from final prediction: effects of revisions - since last forecast only

	(I)		(II)		(III)		(IV)	
	Home	Away	Home	Away	Home	Away	Home	Away
Revisions ($\widehat{\gamma}$)	0.44***	0.35***†††						
	(0.04)	(0.03)						
Revisions, +ve ($\widehat{\gamma}_p$)			0.25	0.07†††	0.25	0.07†††	0.35***	0.03†††
			(0.19)	(0.05)	(0.19)	(0.05)	(0.09)	(0.04)
Revisions, -ve ($\widehat{\gamma}_n$)			0.62***	0.70***†††	0.62***	0.70***†††	0.53***	0.77***†††
			(0.19)	(0.06)	(0.19)	(0.06)	(0.09)	(0.04)
Days to kick-off	0.0182	-0.0207	0.0170	-0.0208	0.0164	-0.0208	-0.0028	-0.101
	(0.022)	(0.012)	(0.022)	(0.012)	(0.022)	(0.012)	(0.009)	(0.007)
Days to ... squared	-0.0001	0.0003	-0.0001	0.0003	-0.0001	0.0003	-0.0001	0.0001
	(0.0003)	(0.0002)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Days taken to revise	-0.0076	-0.0201	-0.0073	-0.0158	-0.0076	-0.0158	-0.0386***	-0.0094
	(0.018)	(0.013)	(0.018)	(0.012)	(0.018)	(0.012)	(0.008)	(0.006)
Days taken ... squared	0.0001	0.0002	0.0001	0.0001	0.0001	0.0002	0.0005***	0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Revision, other ($\widehat{\psi}$)					0.67**	-0.042	-0.01	0.45***
					(0.34)	(0.08)	(0.33)	(0.12)
Round fixed effects	Yes		Yes		Yes		No	
Player fixed effects	Yes		Yes		Yes		No	
Match fixed effects	No		No		No		Yes	
R^2	0.201	0.177	0.201	0.187	0.202	0.187	0.873	0.847
N	1,492	2,093		2,076	2,290		2,076	2,290

***,** indicate significance from one at 1% and 5% levels,, respectively, two-sided tests. Cluster-robust standard errors estimated and displayed in parentheses, with (131/148) clusters at the player level, except for (III), where (380) clusters are at the match level.

For "Revision", †††,†† indicate significance from one half at 1%, and 5% levels, respectively, two-sided tests.

(I)-(IV): dependent variable is goals final forecast error (actual minus forecast)