

# Testing Bayesian Updating with the AP Top 25

Daniel F. Stone\*

Johns Hopkins University

October 2007

[WORK IN PROGRESS—COMMENTS WELCOME.

DO NOT CITE WITHOUT AUTHOR'S PERMISSION.]

## Abstract

Most studies of Bayesian updating are based on experimental data. I use a novel, real world data source—the Associated Press college football poll, a weekly ranking of the top 25 teams—to test the validity of Bayes' rule as a descriptive model. I argue that the voters' final rankings represent their 'true' rankings, for a given season. In the primary analysis I use historical score data and final rank frequencies to estimate Bayesian updated ranks for the individual voters. I compare estimated to actual posterior ranks, and find evidence that the poll voters systematically over and underreact to new information, depending on the circumstance. Overreaction is positively associated with the salience of new information, and lack of salience of strong priors. These results are supported by those found using several alternative methodologies. To some extent the errors nullify each other, causing average observed posterior beliefs to be close to Bayesian posteriors.

**JEL Classification Numbers:** D80, D83, D84

**Keywords:** Bayesian Updating, Overreaction, Underreaction, Salience, College Football Rankings.

---

\*I am very grateful to my advisors, Edi Karni and Matt Shum, for all of their help. I also thank Paul Montella of the Associated Press for providing me with the 2006 ballots and helpful discussion, and Tumenjargal Enkhbayar, Steven Shore, Tiemen Woutersen and Peyton Young for helpful comments. All errors are mine. Address: Daniel Stone, JHU—Department of Economics, 3400 N. Charles St., Baltimore, MD 21218. Email: stone@jhu.edu.

# 1 Introduction

Economic models typically assume that agents update their beliefs rationally, i.e. using Bayes' rule. Currently it is well known that this is a questionable assumption,<sup>1</sup> but it is not clear what, if anything, should be used in its place. This is because most of the evidence of departures from Bayesian updating has been found in the laboratory,<sup>2</sup> and perhaps more importantly, the evidence is varied, and does not suggest one single alternative rule.

The fact that much of the research showing that people do not use Bayes' rule is experimental subjects it to a variety of criticisms regarding its real-world validity. A general objection to the merits of experimental evidence is that in experimental settings agents lack expertise and the ability to learn. Many experiments attempt to address this concern by, e.g., giving subjects opportunities to practice. Still, the issue can never be completely mitigated. The intuition real-world agents gain from years of experience is not replicable in the lab. Other criticisms of experimental results include self-selection of agents, small stakes, self-consciousness of agents, agents not having the ability to confer with others<sup>3</sup> and agents not having sufficient time to make optimal decisions.<sup>4</sup> A perhaps more subtle weakness of experimental research is that it may be relatively unlikely to lead to unexpected findings, since experiments are structured to test particular hypotheses formed a priori.

Regardless, experimental evidence is extremely valuable, and if it conclusively suggested a single alternative to Bayesian updating this would likely be embraced. There is a second significant problem with this research, however, which is that it can appear to be contradictory. The two systematic alternatives to Bayesian updating are overreaction and underreaction to new information. There is substantial research indicating that individuals may do either, depending on the circumstance. People overreact when they make mistakes such as the *representativeness heuristic*, the *base-rate fallacy* and the *hot-hand fallacy*. These biases cause individuals to put too much weight on the signal and/or insufficient weight on the prior. People underreact when they show evidence of, e.g., the *anchoring* and *confirmatory* biases, and make the opposite mistakes regarding the signal and/or prior.<sup>5</sup> Most experimental studies concentrate on showing that one or the other bias or fallacy exists, and are consequently incapable of answering the question of which bias occurs when, and how biases may interact.

This paper attempts to contribute to this research area by addressing both of the concerns described above. I address the issue of the potential weakness of experimental findings by using a novel, real-world data source: the AP college football poll. The AP college football poll is a weekly ranking of

---

<sup>1</sup>See, e.g., Tversky and Kahneman (1974), Grether (1980), and Rabin (1998).

<sup>2</sup>DellaVigna (2007) reviews 'field' evidence of behavioral anomalies, and does not include any discussion of non-Bayesian updating.

<sup>3</sup>Charness, Karni, and Levin (2007) find strong evidence of the significance of this factor.

<sup>4</sup>Levitt and List (2007) provide an interesting discussion of the generalizability of experimental results given these issues.

<sup>5</sup>For a more thorough discussion of these biases see the references noted in footnote 1.

the top teams by 65 experts (journalists who have covered college football for a substantial period of time). Although the data are not economic, they are unique in that they allow us to directly observe the evolution of experts' beliefs over time in response to relatively few clearly observed signals. I discuss the data source in detail in Sections 2 and 3. And I hope to address the question of over versus underreaction through the richness of the data: if deviations from Bayesian updating are found, we may also gain some insight into the underlying causes of the different types of errors.<sup>6</sup>

The primary analysis is of the individual voters' behavior in the 2006 season, which I describe in Section 3. Here I use historical score distributions and empirical final rank frequencies to estimate benchmark Bayesian updated rankings. I then compare these Bayesian estimates to the actual updated rankings and test for systematic deviations. The Bayesian estimates are more accurate than the actuals, in a sense that will be made precise, though not overwhelmingly so. I find evidence that voters do in fact systematically underreact in some situations and overreact in others. The factor that most parsimoniously appears to explain the phenomena is salience: the tendency to overreact increases with the salience, or notability, of the signal and/or prior.

In Section 4 I describe two robustness checks. In the first, I directly compare the actual posteriors to the teams' final ranks. In this case differences can be interpreted as forecast errors rather than non-Bayesian behavior per se. In the second robustness check, I *assume* the voters are Bayesian and estimate the priors that rationalize the observed behavior. I find results supporting the initial conclusions using both of these methods.

Section 5 discusses an analysis of the aggregate polls. These data are less appealing because, obviously, we are unable to track individual behavior as closely. Still, they are valuable as they allow us to examine different seasons and beliefs over a wider range of ranks (as will be explained), and there are much more extensive historical data. The findings from this section support our initial conclusions as well. On average, however, the actual aggregate ranks outperform the Bayesian estimates by the accuracy metric used.

## 1.1 Closely Related Literature

While, as alluded to, the literature on behavioral belief updating is vast, there are a few papers in particular that stand out as being closely related to this study. Amir and Ganzach (1998) analyze over and underreaction using field data (stock analysts' earnings forecasts). While their sample is large, their data are much less rich in that they do not include data on the signals that cause beliefs about future earnings to change. As a result the authors are unable to say anything about what the theoretically correct Bayesian responses to the signals are. In lieu of this they focus their analysis on

---

<sup>6</sup>There is a substantial literature on over and underreaction in financial markets (e.g. DeBondt and Thaler (1985) and Hong and Stein (1999)). Most of this work does not apply to the issue at hand, belief updating at the micro level in a non-competitive setting.

earnings forecast errors. They are able to show that the errors systematically vary depending on the type of forecast adjustment (whether it is a mere revision, or a more drastic change). Their data are highly confounded by optimism (which they clearly recognize) and hot-hand biases (which they do not). The authors' findings that saliency and overreaction are positively related are consistent with, but more limited than, those discussed in this paper.

Offerman and Sonnemans (2004) attempt to test competing explanations of overreaction, motivated by the numerous empirical studies documenting the phenomenon in financial and sports betting markets. The two hypotheses they compare are *recency* and the *hot-hand fallacy*. The former is something of a catch-all for mistakes involving over-weighting recent information, while the latter depends on agents mistakenly inferring positive serial correlation, or trend, in the unobserved state.

The authors design an experiment to distinguish between these possibilities. A coin, which with 50-50 probability is either fair (and i.i.d.) or serially correlated, is tossed 20 times. Participants observe the 20 results and are then asked to estimate the probability the coin is correlated. The authors claim the hot-hand fallacy better explains their data. One issue that could be raised with their study is that participants do not respond to single signals, but a large string of signals. First of all this makes the task unrealistically complicated (although they find performance does not improve substantially with training). Second, seeing a string of signals may make participants more likely to see patterns, or false correlation, than a setting in which responses can be made after any number of signals.<sup>7</sup> In addition, the authors do not address situations in which the hot-hand is not applicable (it is known—or should be known—that there is no trend in the underlying state). Their study also precludes analysis of underreaction and other causes of overreaction.

There are several economic studies that use the AP college football poll as a data source. Goff (1996) uses the poll to analyze path dependence in collective decision-making. He finds that final poll outcomes are affected by preseason polls while controlling for team performance. This is completely consistent with my approach, as I assume the preseason poll incorporates valuable information that may also impact final beliefs. Logan (2007) tests several college football hypotheses relating specifically to the poll. Among his findings are that the voters' responses to late season losses are smaller than those to early season losses. This is also consistent with my approach since, in a belief updating framework, adjustments to beliefs grow smaller as more information is accumulated (i.e. throughout the season). Lebovic and Sigelman (2001) is most similar in spirit to this paper. Their findings are similar to mine, however, they are less precise as the authors do not explicitly test Bayesian updating and do not use data on individual voter behavior.

---

<sup>7</sup>It would be interesting to see how their results change if respondents had an incentive to guess whether the coin was fair or not as early, i.e. after as few flips, as possible.

## 2 The Data

The AP college football poll is a data source uniquely suited for analyzing belief updating. This is because, during the season, the poll is conducted exactly once per week and teams play exactly one or zero games per week. Consequently, the voters in the poll observe at most one major signal about each team per week. (Admittedly the voters observe other information about each team besides game scores every week, but this information is relatively insignificant, especially on a week-to-week basis.) Moreover, the signal probabilities—the distributions of the scores—are common knowledge, since the voters have all observed years of scores. Based on their extensive experience the voters should know how likely different scores are for teams of different ranks.

These two features—the single signal between observations of the voters’ rankings, and common knowledge of the signal distributions—are what distinguish these data from most economic data, and are why I use them for this study. In most economic situations there are many important signals, which arrive erratically, that may affect beliefs. It is difficult to tell which individuals observe which signals, and even more difficult to say anything about the (subjective) likelihoods of the signals. For example, take the individual’s lifetime consumption-savings problem, which depends on beliefs about future income and other factors. Even if we observe an individual’s history of income and consumption we can really say very little about the signals she has received about future income (including personal data, such as how much her boss at her current job likes her at the moment, and the individual’s knowledge of the possibly relevant current account deficit). Thus we cannot say, based on observed data, whether or not the individual uses Bayes’ rule to update beliefs upon receipt of new information.

The aggregate AP poll is widely available. The poll is conducted weekly throughout the season, and once both before it starts and after it ends. The season begins in late August and ends in early January. The poll is voted on by approximately 65 leading college football journalists from throughout the country and different forms of media. Each voter submits a ranking of the top 25 teams, and the aggregate ranking is determined by assigning teams 25 points for each first place vote, 24 for second, etc., and summing points by team (a *Borda* ranking). The poll began in 1934 but the number of teams ranked by voters has changed over time, and has been 25 since 1989. Historically, the poll has played a part in determining the national championship, however, this role ended in 2005. The individual ballots of the AP poll voters are not confidential. The AP currently makes the current week’s ballots available on its website, but the historical ones are not published anywhere to my knowledge. I obtained historical aggregate AP polls and ‘Others Receiving Votes’ (teams receiving some votes whose point totals were not in the top 25) from [appollarchive.com](http://appollarchive.com) and the Baltimore Sun. I obtained the individual ballots for the 2006 season from Paul Montella and Ralph Russo of the Associated Press. Historical score data is from <http://homepages.cae.wisc.edu/~dwilson/rsfc/history/howell/> and <http://www.knology.net/~jashburn/football/archive/>.

The data do have several weaknesses. First, the voters do not have direct incentives relating to the qualities of their rankings. This is not too concerning, as the voters' prestige, and thus indirectly career concerns, depend on their rankings. For example, a voter was removed from the 2006 poll after mistaking a win for a loss (<http://sports.espn.go.com/ncf/news/story?id=2663882>), and another voter is famous for being the only one to rank the eventual championship winner #1 early in, and consistently throughout, the 1992 season. Moreover, discussions with voters indicate that they put substantial effort into producing their best possible rankings.

There are two more significant weaknesses. The first is that the voters only rank 25 out of more than 100 teams. The second is that we only observe rankings, rather than specific beliefs regarding a particular variable. What the voters are ranking is never formally defined. Likely, there are different criteria used by different voters. I discuss both of these weaknesses at length in the following section.

### 3 The Individual Voters

In this section I discuss explicitly estimating the benchmark Bayesian updated rankings for the individual voters and results obtained when comparing the Bayesian estimates to the observed data.

#### 3.1 Defining *Truth* and an Estimation Framework

The key element to estimating the Bayesian ranking changes is the definition of *truth*—the unobserved state of the world that the rankings represent beliefs about. A definition of truth is needed so that we can estimate the conditional signal distributions (the likelihoods of the game scores conditional on true rankings) and the prior distributions (the likelihoods of true rankings before the games). The rankings are never formally defined so the definition of truth is not obvious. Some definitions imply that the rankings are completely subjective and prohibit the econometrician from saying anything about rational belief changes given game results. The most plausible example of a ranking definition for which we would have this problem is assessment of year-to-date (YTD) performance. If voters ranked teams according to subjective assessments of YTD performance game scores would not be signals relating to truth; they would in fact be truth. I do not believe this ranking definition, however, is plausible. This is because of the existence of a preseason poll. Since there is no YTD performance at that point, and a poll exists, the poll cannot be an assessment of performance that has been observed. Moreover, polls from early in the season are based on minimal actual performance, and so cannot sensibly be interpreted as based of YTD performance.<sup>8</sup>

---

<sup>8</sup>In an email one voter stated, 'Starting with week 1, I base my poll on body of work only.' The term 'body of work' appeared to refer to YTD performance. The voter's rankings, however, were clearly at odds with this statement. For example, the voter ranked USC in the top 5 and Nebraska in the 20-25 range in the preseason poll. They both had fairly strong wins over historically weak teams in week 1 (Nebraska won by a slightly greater margin), thus their YTD performances at that point were very similar. But the voter's rankings of the teams did not change significantly in the

I believe there are only three plausible definitions of the rankings, which I denote D1, D2 and D3: (D1) an ordering of some statistic (e.g. expectation or median) of beliefs about current, unobserved quality; (D2) an ordering of quality of season-long performance, which is unobserved in all polls except the final one; and (D3) a combination of D1 and D2. I intentionally leave the terms ‘quality’ and ‘quality of season performance’ undefined to avoid imposing unnecessary structure. For concreteness, it may be useful to think of quality as relating to likelihood of victory; if team A’s quality is greater than B’s than team A’s probability of beating B is greater than 0.5.

I argue that for each ranking definition the voters’ individual final rankings are optimal estimators of truth. While we may never observe truth we can estimate it, and this is sufficient to conduct the subsequent analysis.

Formally, D1 can be represented as follows. Let  $X_{i,t}^v$  denote a team’s subjective unobserved true quality in week  $t$  for voter  $v$ . Quality is subjective because I do not restrict its definition to be constant across voters. It is subscripted by  $t$  because it may change over time. Let  $r_{i,t}^v$  denote the true rank of team  $i$  in week  $t$  for the voter  $v$ .  $v$  is henceforth suppressed. The true rank is also subjective—it is simply the rank ordering of true qualities:

$$\text{D1} : X_{i,t} < X_{j,t} \rightarrow r_{i,t} > r_{j,t}. \quad (1)$$

Since  $X$  is never observed the true rankings are never observed. Suppose we assume  $X_{i,t} = X_i$  for all  $t$  and  $i$ , i.e. quality is constant for all teams. This assumption may appear questionable. We will return to discuss its validity shortly. If we make the assumption, though, it is clear that  $r_{i,t} = r_i$  for all  $i, t$ , and the variance of beliefs about  $X_i$  will decrease throughout the season as information is acquired about  $X_i$ , i.e.  $\text{Var}_t(X_i) \leq \text{Var}_{t'}(X_i)$  for all  $t' < t$ . Consequently the voters’ actual rankings from the final period (the postseason rankings) will have lower variance than their rankings from any other week of the season. Let  $\bar{t}$  denote the postseason week and  $\tilde{r}_{i,t}$  denote a voter’s actual rank of team  $i$  in week  $t$ , i.e. the rank the econometrician observes. Then  $\tilde{r}_{i,t}$  is an optimal estimator of truth for each voter-season; it has minimal variance and is unbiased.

Now we return to D2, the interpretation in which voters rank teams on quality of season performance. By definition  $r_{i,t}^v = r_i^v$  for all  $i, t$  and  $v$ , since the true rankings are based on a single measure for the entire season. Let  $Y_{i,t}^v$  denote a voter’s subjective belief about team  $i$ ’s quality of performance in week  $t$ . Again  $v$  will be suppressed whenever possible. Let  $G_i$  denote the number of games team  $i$  plays in the season. Then<sup>9</sup>

$$\text{D2} : \frac{1}{G_i} [\sum_{t=1}^{G_i} Y_{i,t}] < \frac{1}{G_j} [\sum_{t=1}^{G_j} Y_{j,t}] \rightarrow r_i > r_j. \quad (2)$$

---

following poll. Thus the rankings could not be based only on YTD performance starting in week 1.

<sup>9</sup>This definition assumes each game of the season is weighted equally for simplicity but without loss of generality.

In this case the variable on which the true rankings are based,  $\sum_{t=1}^{G_i} Y_{i,t}$ , is observed by the end of the season (week  $\bar{t}$ ). Thus  $\tilde{r}_{i,t} = r_i$  when voters use D2, i.e. the actual postseason rankings are in fact truth in this case and are thus trivially also optimal estimators of truth. Since the postseason rankings provide the best estimate of truth for both D1 and D2, and D3 is a ranking based on a combination of these measures, the postseason rankings clearly also provide the best estimate of truth for D3.

To sum, independent of the voter’s preferred ranking interpretation, the individual voters’ post-season rankings are optimal estimators of truth, given the constant team quality assumption. Perhaps surprisingly, the data show that this assumption is not unreasonable. This may be because of the relatively short season (teams play fewer than 15 games)—or may be additional evidence of the lack of a hot hand at the team level Camerer (1989)). I test this assumption by testing the hypothesis that average score differences for teams of different final ranks are constant throughout the season. If voters rank teams on current quality and quality changes throughout the season, we would observe relatively better performances in the later part of the season by teams relatively highly ranked in the final poll.<sup>10</sup> This is because in this case teams highly ranked in the final poll improve on average throughout the year, and teams ranked poorly in the final poll worsen. Please see the estimation appendix for a brief example illustrating this phenomenon, which, if it were to occur, would cause us to reject the hypothesis that average score differences are constant throughout the season.

Table 1 presents empirical evidence indicating that the hypothesis that score differences are constant over time should not be rejected. While home teams of final rank 1-12 do beat teams of final rank 13-25 by a greater margin in the second half of the season, we also see that home teams of final rank 13-25 perform better in later months versus superior teams of rank 1-12. These data essentially nullify each other. Neither of the other p-values are compelling either. We have little reason to lose confidence in the null. This allows us to use the voters’ final rankings to estimate the true rankings.

I note that truth as defined here is endogenous: it is determined by the voters. Consequently, their belief updating from the second-to-last to last poll is tautologically Bayesian. In order to test Bayesian behavior then we must examine earlier polls—in fact, the earlier the poll, the greater the statistical power will be for the tests. On the other hand, if we restrict the sample to too small a portion of the season we also lose power. I thus limit the analysis simply to the first half (seven weeks) of the season.

With this discussion in mind we can now specify the voters’ objective functions and Bayesian updating process. Let  $r_i^v$  continue to denote true rank of team  $i$  for voter  $v$ ,  $i \in \{1, \dots, 120\}$ ,  $r_i \in \{1, \dots, 25, 26+\}$  and  $v \in \{1, \dots, 64\}$ , in which if  $r_i = 26+$  the team is unranked.  $v$  is again suppressed in the following as it is unnecessary. Note that  $i$  is a team identifier, or index, and has nothing to do with the team’s rank. For example,  $r_1$  may be 10, 20, etc.

Each voter’s objective function in week  $t$  is to minimize a function of distance between current and

---

<sup>10</sup>If voters rank teams on season performance then we would not observe a relationship between performance and time. But in this case the constant quality assumption is not needed to use final rank as an estimator of true rank.

true ranks. I assume they minimize a sum of quadratic loss functions:

$$\tilde{r}_t = \underset{r_t}{\operatorname{argmin}} E_t[\sum_i (r_{i,t} - r_i)^2], \quad (3)$$

in which  $\tilde{r}_t$  is the vector of ranks chosen in week  $t$ . The FOC imply  $\tilde{r}_{i,t} = E_t(r_i) \forall i, t$ . However,  $E_t(r_i)$  is non-integer while  $\tilde{r}_{i,t}$  is constrained to be in the integer set noted above that  $r_i$  is an element of. To reconcile these things I assume  $E_t(r_i) > E_t(r_j) \rightarrow \tilde{r}_{i,t} > \tilde{r}_{j,t}$ ; teams are ranked in order of expected rank.

Let  $s_{ij}$  be a random variable that is the score difference from the game in which team  $i$  plays team  $j$ . That is, it is points scored by team  $i$  minus points scored by  $j$  (if  $s_{ij} > 0$ ,  $i$  wins). This variable has no time subscript because teams almost never play each other more than once.

Let  $g(s_{ij}|r_i, r_j)$  be the conditional probability that the game between teams  $i$  and  $j$  results in score  $s_{ij}$ , given their true ranks.

Let  $f_{i,t}(r_i)$  be the (subjective) probability that team  $i$  has true rank  $r_i$  in week  $t$ .

After team  $i$  plays  $j$ ,  $s_{ij}$  is observed and voters can update their beliefs to  $f_{i,t+1}(r|s_{ij})$ ,  $f_{j,t+1}(r|s_{ij})$ . I note that technically if beliefs about team  $i$ 's rank change, beliefs about at least one other team's rank also change. I.e.  $\forall k \neq i, j$  the voters may update  $f_{k,t+1}(r|s_{ij})$ , but since these effects are minimal I abstract from them. Similarly, I assume  $f_{i,t}(r_i|r_j) = f_{i,t}(r_i), \forall j \neq i$ .

Voters know  $g(s_{ij}|r_i, r_j)$  from their observation of years of historical scores and estimates of the true rankings (the postseason rankings) for the respective seasons. They can thus use a fairly straightforward application of Bayes' rule to update beliefs. For example, suppose teams indexed 10 and 11 play a game and we are interested in the posterior probability that team 10 has true rank 1:  $f_{10,t+1}(1|s_{10,11})$ . This is the probability of  $s_{10,11}$  given  $r_{10} = 1$ ,  $g(s_{10,11}|r_{10} = 1)$ , times the prior that team 10 has true rank 1,  $f_{10,t}(1)$ , divided by the unconditional score probability,  $g(s_{10,11})$ . This last term depends on beliefs about the true ranks of teams 10 and 11, specifically  $g(s_{10,11}) = \sum_{r_{10}} [\sum_{r_{11}} g(s_{10,11}|r_{10}, r_{11}) f_{11,t}(r_{11})] f_{10,t}(r_{10})$ . In general, the formula for belief updating is:

$$f_{i,t+1}(r_i|s_{ij}) = \frac{g(s_{ij}|r_i) f_{i,t}(r_i)}{g(s_{ij})} = \frac{[\sum_{r_j} g(s_{ij}|r_i, r_j) f_{j,t}(r_j)] f_{i,t}(r_i)}{\sum_{r_i} [\sum_{r_j} g(s_{ij}|r_i, r_j) f_{j,t}(r_j)] f_{i,t}(r_i)}. \quad (4)$$

Note that  $g(s_{ij}|r_i)$  is calculated by averaging over  $r_j$ , the true rank of team  $j$ , since this is unobserved, and likewise for  $g(s_{ij})$ . The econometrician can also then estimate posterior beliefs if the prior and signal distributions are estimable, which is the subject of the next subsection. If we can translate these estimated posterior beliefs into rankings we can compare estimated Bayesian posterior rankings to actuals.

One issue with this framework that should be noted is that it assumes voters have no one to please

but themselves; their objective functions do not depend in any way on others' perceptions of the accuracy of their rankings.<sup>11</sup> This is not accurate if voters are inhibited from expressing their views for fear of embarrassment, or due to other factors. An alternative estimator of truth that would account for these potential interaction effects would be the aggregate postseason rankings. This estimator is less appealing because it does not allow for subjectivity of final beliefs. Still, I will estimate Bayesian posteriors using this alternative truth estimator in the next version of this paper, and compare them to the original estimates. I do not expect the differences to be material.

## 3.2 Estimation Method

This subsection describes estimation methodology and can be skipped without any loss of continuity. I make many assumptions, some of which may appear fairly arbitrary. The resulting estimates would be suspect if we could not assess their validity and/or check sensitivity to the assumptions they are based on. Fortunately I am able to do both of these things, which I discuss in other parts of the paper.

### 3.2.1 Distributions

To estimate the Bayesian posteriors we first need estimators of both of the components of 3.1, the  $f$ 's and  $g$ 's. I start with the  $g$ 's—the signal distributions. The natural way to estimate these is to use the historical distributions of scores between teams of the various final ranks. Because I do not have historical individual final rank data, I need to use the aggregate final rank data for this purpose. I do not believe this substitution introduces significant error, especially since there is a large degree of convergence in beliefs in the final poll. The assumption is equivalent to assuming each voter uses the same score distributions for belief updating, and these distributions are the distributions conditional on aggregate rank. If anything, this assumption may cause the estimated score distributions to be less noisy than they really are. This will cause the signals to appear too informative, or bias the estimates towards overreaction. We will keep this in mind as we proceed.

The other issue with this component is that although I have access to all historical scores, the sample sizes for scores between teams of particular ranks is highly limited. Recall that I use score data dating back to 1989 because that is when the top 25 AP poll in its current form began.<sup>12</sup> In 17 years of data there are very few games between teams of each rank combination during the regular season since it is so short. For example, there were exactly two games between teams of final rank 1 and 2 played during the regular season from 1989-2006. In addition, I condition the distributions on home/away

---

<sup>11</sup>I do test for, and find significant, the effects of differences between individual and aggregate ranks on individual rank changes. I believe this is likely to be due to learning—voters know that other voters have superior information about some teams, and they justly influence one another through their rankings. I discuss this later in the paper.

<sup>12</sup>I use data from all regular season games but exclude games played at neutral sites. I do not exclude any games due to injuries. The significance of injuries in the sport is very difficult to determine—many teams have had very good seasons with multiple seemingly major injuries (e.g. Nebraska 1994, Louisville 2006). I believe attempting to clean the data this way would create more noise than it would eliminate.

status, to account for this variable affecting the distributions in different ways for teams of different ranks. As a result I am forced to use multiple smoothing techniques. First, I break the ranked teams into four groups; 1-6, 7-12, 13-18, and 19-25. This categorization is the finest that yielded relatively large sample sizes ( $n > 20$ ) for games between teams in each sub-group. Then, I discretize the score distribution into buckets of size 7 (with upper and lower bounds of plus/minus 49+), and use moving averages to smooth the empirical frequencies in these categories. An example of the unsmoothed and smoothed histograms is given in Figure 1. More detailed description of this method is in the estimation appendix.

Estimating the  $f$ 's—the priors—is actually less straightforward. One might naturally assume here that we would estimate different priors for each voter and team. The available data, however, are not nearly extensive enough for this task. The next thought might be to parameterize beliefs about teams as functions of observable team characteristics and then somehow randomize them over the individual voters to create heterogeneity. A simpler, less data-intensive and less parametric method is to estimate the distributions by prior *rank*, rather than team. Then, if we make the plausible assumption that all voters have the same beliefs, or probabilities of the various true ranks for teams of the same prior rank, we only need to estimate as many prior distributions as we have prior ranks, for each week.

**Example 3.1.** *To illustrate, suppose voters A and B have UVA and UNC ranked (1,2) and (2,1) respectively in week one. They cannot both have the same prior beliefs for both teams since they rank the teams differently. Rather than estimate priors for UVA and UNC by voter (four distributions), we can estimate the priors for (generic) teams ranked 1 and 2 (two distributions). The cost of this simplification is that it implies voter 1 has the same prior for its # 1 team (UVA) that voter 2 has for its #1 team (UNC). Later in the paper I discuss testing this assumption.*

Using this approach it is then natural to estimate the prior distributions using the empirical frequencies of our estimator of true rank (final rank) conditioned on current rank. Recall that by assumption the voters have rational expectations:  $\tilde{r}_{i,t} < \tilde{r}_{j,t} \rightarrow E_t(r_i) \leq E_t(r_j), \forall i, j, t$ . In Table 2 we see that due to the limited sample size, monotonicity of expected rank is violated in the raw frequencies.<sup>13</sup>

To account for this inconsistency I again apply smoothing techniques to the distributions. In this case the key variation is between, rather than within, initial ranks, and I smooth the distributions accordingly. That is, for each prior rank, I estimate the probability of finishing in a rank group as the empirical probability of finishing in the group for teams with the prior rank and neighboring higher and lower prior ranks. An illustrative example would be to estimate the probability a team of prior rank 2 finishing 1-6 as the frequency of teams with prior ranks 1-3 finishing 1-6. This method preserves the appropriate features of the joint distribution of rankings, which I refer to as ‘rank accounting’

---

<sup>13</sup>I calculate expected rank by computing the weighted average of midpoint ranks from each rank group and 35 for the unranked category. Results are not sensitive to the choice of value for unranked teams.

constraints; e.g. that the expected number of teams finishing 1-6 is six. The practical question then is how many neighboring ranks should I use for this purpose? An extreme approach would be to use the minimum number to obtain monotonicity of expected rank over all 25 prior ranks. The problem with this is that it requires over 5 neighbors for each week (both above and below), which can lead to over-smoothing for some prior ranks. In general, we obtain monotonicity for teams with better initial ranks using fewer neighbors than needed for worse initial ranks. If we used the higher number of neighbors for all teams we would over-smooth the distributions for the better ranked teams, which would bias the estimates of the priors for those teams (estimated expected rank would be worse than it should be). To balance these considerations I use a mid-point threshold to determine the number of neighbors to use: the minimum number that yields monotonically increasing expected rank for at least 13 consecutive prior ranks. Rather than use a different number of neighbors for the other teams, I simply average the frequencies when monotonicity is violated. When a high number of neighbors is required, the distributions are so similar that assuming them to be equal is a reasonable approximation. Please see the estimation appendix for more detailed discussion of this procedure.<sup>14</sup>

I note that it is somewhat disconcerting to use the 2006 data to estimate the priors for the 2006 season. Ideally, we would use historic data for this purpose. However, this method does not necessarily bias the analysis, as the actual final rank frequencies for any given season are unbiased estimators of the prior beliefs about final ranks for that season. Estimating the priors using same-season data does not imply the voters have perfect foresight, only that they use the ‘correct’ priors for the season. If there is some uncertainty about what the correct priors are, however, the estimated priors may be too ‘strong’, or voters would be estimated to be too confident in their beliefs. This would bias estimated reactions towards zero, indicating underreaction. Recall that use of aggregate rankings for estimation of the score distributions was noted to bias estimated reactions, if at all, away from zero. Hence, the two potential biases point in opposite directions. Since neither one is likely to be significant on its own, we conclude that the bias of the estimates overall should be minimal.

### 3.2.2 Limited Rankings

Another important weakness of the data is that the voters only rank 25 out of greater than 100 teams. We do not observe individual voters’ beliefs for any of the other teams, although the voters surely do not consider all unranked teams to be equal. However, since most games are between ranked and unranked teams, we need some objective method of differentiating among unranked teams. If we treated wins over the best and worst unranked teams equally our estimates would be badly flawed.

Fortunately, there are a number of publicly observable variables that we can use to differentiate

---

<sup>14</sup>I ‘wrap around’ neighbors at ranks 1 and 25, double-counting ranks 1 and 25 to preserve rank accounting; e.g. a one-neighbor average frequency for prior rank 1 would be the average frequency of teams with prior rank 1, counted twice, and prior rank 2. The mid-point threshold method discussed above does not satisfy rank accounting constraints precisely, but the errors are minimal.

unranked teams. I use three: 1) currently ranked by at least one other voter, 2) ranked by at least one voter in final AP poll in one of previous two seasons, and 3) ranked by at least one voter in final AP poll in one of previous three to five seasons (and unranked in previous two seasons). I also distinguish by YTD number of losses (0 or  $>0$  in weeks 1-3; 1 or  $>1$  in weeks 4+) for teams not currently receiving votes from another voter. This expands the size of the set of elements  $r_{i,t}$  is in to 32, in which  $r_{i,t} = 26$  means team  $i$  receives votes from others,  $r_{i,t} = 27$  means team  $i$  does not receive any votes but was ranked in one of two previous seasons and has zero losses, etc. I estimate priors for teams in each of these groups as the raw frequencies of finishing in the various rank categories (I do not smooth them because there is no a priori criterion for smoothing as above for top 25 teams, since specific beliefs on the teams are not observed. But, generally teams receiving votes from others, and teams ranked in recent polls fare better than others.).

This method of distinguishing among unranked teams is not sufficient for accurately estimating posterior beliefs for unranked teams. Consequently, I only estimate posterior beliefs for teams that are currently ranked. I do not attempt to estimate which teams should enter and exit the top 25. This forces a need to account for the fact that several teams do indeed drop from the rankings for most voters in most weeks. I do this by restricting the maximum (worst) estimated posterior rank to one greater than the number of teams that are observed to stay in the poll, by voter-week. I also re-rank observed posterior ranks among teams that were in the prior poll, and assign the same maximum rank to teams that drop from the rankings. This allows comparisons between estimated and actual posteriors to be apples-to-apples, unconfounded by teams entering the polls at various rank levels.

**Example 3.2.** *This method is again best illustrated by example. Suppose only 22 of 25 teams in voter 1's week 1 ballot are ranked in week 2. Suppose the teams ranked 19-21 in week 1 dropped out and were replaced by new teams (teams unranked in week 1), so the ranks of teams ranked 1-18 and 22-25 did not change. Since I know nothing about the new teams in the poll I ignore them and adjust the actual week 2 posteriors. I assign ranks 19-22 to teams actually ranked 22-25, and 23 to the teams that dropped out. For the estimates, I assign rank 23 to all teams with estimated rank 23 or higher. This method allows the estimated posteriors to potentially exactly match the actuals.*

### 3.3 Results

#### 3.3.1 Validity of the Estimates

Before comparing estimated Bayesian ranking changes (posteriors) to actual (observed) ranking changes, it is worthwhile to attempt to assess the validity of the estimates. I do this by examining the distances between the estimated true rankings and the estimated posteriors. I compare these to the same distances from truth for the actual posteriors, actual priors and flat priors. Theoretically, the Bayesian

rankings are on average closer, by any reasonable metric, to truth than posterior rankings obtained using any other method. Thus, if the estimates and actuals are both approximately Bayesian they should be equally close in distance to the true rankings. If either is less Bayesian its distance from truth will be greater. If either the estimates or actuals is sufficiently flawed and it will be no closer to truth than the actual priors. If the signal is uninformative neither the estimates nor actuals will be closer to truth than the actual priors. If the actual priors are meaningless they will be no closer to truth than flat priors (uniform prior beliefs).

I measure distance from truth by average absolute deviation:  $\frac{1}{n} \sum_{i,t} |\tilde{r}_{i,t+1}^v - r_i^v|$  ( $n$  is the number of observations). I adjust the estimated true rankings as discussed above to account for number of teams, per week and voter, not being in the final poll. Table 3 depicts summary statistics for these deviations averaged over all voters. We see that the estimates appear superior to the actuals, and that the signal and actual priors appear informative, as they are both lower than the flat prior distance average deviation. We cannot formally test these differences here, however, because the observations are not i.i.d.; many of the observations are correlated by game and voter. Instead, I conduct paired t-tests by individual voter of the difference between the average absolute deviation from truth for estimated and actual posteriors. The null is that the averages are the same, and the alternative is that the average deviation for the estimates is smaller. The average p-value (for the 64 voters) is 16.7%, with a min of 0.04% and a max of 73.7%. There were p-values below 5% for 37.5% of the voters. Given that the actual estimates incorporate information and personal biases not accounted for by the estimates, I interpret this as evidence of the validity of the estimates.

### 3.3.2 Testing Bayesian Updating

Although the statistics reported above are evidence that many individual voters are not Bayesian on average, they do not indicate whether or not these voters have systematic biases, and if so, how and when the biases come about. Specifically, I am interested in testing for systematic over and/or underreaction. As mentioned in the introduction there are many well known biases that fall into one of the two categories. I define overreaction formally momentarily, but first discuss summary statistics for the estimated and actual rank changes. These are presented in Table 4, categorized by the basic categories of signal type: win, loss and bye (no game). The estimated and actual responses are similar for both wins and losses, but for wins the magnitude of the average estimated change is greater than the actual, and vice versa for losses. Since we expect ranks to improve after wins and worsen after losses this indicates slight underreaction to wins and overreaction to losses. The estimated and actual responses are markedly different for byes. On average, voters improve teams' ranks after byes, while the estimated ranks worsen. This is interesting since a bye is, itself, no signal. But we can interpret this easily since most games played by top 25 teams are against non-top 25 opponents, and so ranked

teams usually win. Winning is a positive signal, thus the expected rank of most teams that play games improves. Since the expected rank of teams with byes does not change, the rank order (of expected rank) on average worsens. Presumably the voters' beliefs about teams with byes also do not change significantly, thus, since these teams' rankings do not worsen this implies the voters' beliefs about the other, winning teams do not change significantly. This is additional evidence of underreaction to wins.

Rank change ( $r_{\Delta}^i = r_{t-1} - r_t^i$ ,  $i \in \{A, E\}$ ,  $A$ =actual,  $E$ =estimate) is defined so as to represent rank improvement. Note that  $r_{t-1}$  is always the actual prior rank. To define the main variable of interest, overreaction, I transform the estimated and actual rank changes as follows.

$$\text{Definition 3.3. } \textit{Overreaction (OVER)} = \begin{cases} |r_{\Delta}^A| - |r_{\Delta}^E| & \text{if } r_{\Delta}^A * r_{\Delta}^E \geq 0, \\ r_{\Delta}^A - r_{\Delta}^E & \text{otherwise and win,} \\ r_{\Delta}^E - r_{\Delta}^A & \text{otherwise.} \end{cases}$$

A sensible, but cruder definition of this variable would omit the first condition. That is, for 'good' signals (wins) overreaction would be defined as excess rank improvement, and for 'bad' signals (losses), the excess rank deterioration. I believe Definition 3.1 is preferable because it is not true that all wins are good, and losses bad. For example, if a top ranked team struggled to beat a team thought to be of very poor quality, this would probably be interpreted as a bad signal, i.e. the posterior expected mean rank would be worse than the prior. In this case, then, overreaction is the excess rank deterioration. Case 1 of the definition captures this effect. It should be noted, though, that the implicit assumption made is that if either  $r_{\Delta}^A$  or  $r_{\Delta}^E$  is zero, the direction the signal normatively takes is determined by the sign of the non-zero rank change. Using this definition the mean of OVER for wins, losses and byes is -0.341, 0.375 and -2.560.<sup>15</sup>

These numbers are slightly sharper than those that would be inferred from the raw rank changes reported in Table 4. They confirm that voters seem to both over and underreact, and to different degrees depending on the situation.<sup>16</sup> Barberis and Thaler (2002), without citing any studies, attempt to provide an explanation for this apparent contradiction saying, "If a data sample (signal) is representative of an underlying model, then people overweight the data. However, if the data is not representative of any salient model, people react too little to the data." Amir and Ganzach (1998) find some supporting evidence for this idea, but focus on saliency of the 'anchor' (prior) rather than signal.

To attempt to shed additional light on this problem, and control for factors potentially confounding

<sup>15</sup>Results are qualitatively robust to using different definitions of overreaction, but are not reported in the interest of brevity.

<sup>16</sup>One alternative explanation for the data would be that voters put less effort into responding to wins, since they are more common, and that their reactions to losses are not overreactions, but simply more accurate reactions. The data show this not to be the case: the mean deviations from truth are 3.83 and 3.67 for actual and estimated posteriors after wins, and 2.27 and 1.93 after losses, i.e. the voters' posteriors are relatively more accurate compared to the estimates after wins.

the summary statistics, I estimate the following regressions separately for games in which the ranked team wins and loses:

$$OVER_{ij} = X_{ij}\beta + \delta_j + WEEK * \delta_j + \epsilon_{ij}. \quad (5)$$

$i$  and  $j$  subscript game and voter respectively.  $X$  is a vector of controls including the following:

- 1) HOME: dummy for home game;
- 2) WEEK: week of the season;
- 3) SDEV: score deviation from expectation = difference between actual and expected score margin;<sup>17</sup>
- 4) TOP10: dummy for team ranked in voter's top 10;
- 5) OPP25: dummy for opponent being in the voter's top 25;
- 6) SDEV\*OPP25;
- 7) APDEV: aggregate AP rank - voter's rank;
- 8) RKSD: standard deviation of rank by team over voters assuming rank of 35 for unranked teams;
- 9, 10) ST, REG: dummies for the school being in the voter's state or census region (9);
- 11) YRS: years of experience on the poll since 1998;
- 12) ACCURACY: the mean difference between prior and final rank by voter. I use this variable to control for legitimate differences in priors—if some voters' priors are more accurate, either because their priors incorporate more information or other reasons, we would expect them to react less strongly to the signals.

While the definitions of most of these variables are straightforward, I explain their interpretations below. Two other important variables are  $\delta_j$ , a voter fixed effect, and an interaction of this fixed effect with WEEK. This is intended to account for possible heterogeneity of ranking interpretation: although we have established the postseason rankings are equivalent to truth for either individual interpretation, if voters have different interpretations their responses to games might vary. Specifically we would expect D2 voters to react less to signals as the season progresses, which this interaction term should account for. However, we do not expect this difference to be significant since most of the important games occur late in the season and thus even full-season voters should be adjusting rankings primarily due to changes in beliefs about quality.

Results, excluding terms with voter fixed effects, are presented in Table 5 with robust standard errors clustered by game (they are much lower when not clustered or clustered by voter). Given the null hypothesis of Bayesian updating, the expected coefficients for most of the variables is zero. Non-zero

---

<sup>17</sup>Expected score margin is computed using the estimated prior and score distributions.

coefficient estimates are consistent with Bayesian updating for the variables APDEV, RKSD, ACC (and possibly WEEK) because these variables may affect beliefs under rationality. The results are remarkable: many of the variables with expected coefficients of zero under the null are significant at standard levels.

HOME is positive for wins, negative for losses and significant for both. This means voters tend to overreact to home wins and away losses—they do not appreciate the importance of home-field advantage. SDEV is negative and significant for both estimates. Recall that SDEV will be farther from zero for more impressive wins and worse losses, and positive for wins and negative for losses in general. Thus voters are insensitive to margin of victory, but sensitive to margin of loss. Voters react (significantly) more strongly to victory, and margin of victory, when the opponent is ranked. These variables have no significant effect on loss responses. TOP10 is significant for both wins and losses. I believe it is negative for wins due to an interaction with RKSD (they are negatively correlated). The positive coefficient for TOP10 in the losses model is perhaps the most striking of all of the estimates.

APDEV takes the expected signs and is significant at the 1% level for both estimates. Voters are influenced by their peers, either due to learning, herding, or both. RKSD is significant only for wins, and is negative. This indicates voters react less strongly to positive signals when there is substantial disagreement among their peers.

ST, REG and YRS are significant for wins only. The first two are positive, indicating voters react more to wins by teams in their home areas, possibly due to ‘home bias’. The sign on experience (YRS) is negative, indicating voters who have worked on the poll longer make smaller adjustments after wins. The sign is actually positive for losses. Given that we know voters, on average, underreact to wins and overreact to losses, these estimates imply these trends are exacerbated by experience. In other words, experience makes voters more prone to biases—perhaps this is because more experienced voters are more secure about their reputations, and expend less effort in determining rank changes. Neither ACCURACY estimate is significant at conventional levels, but the magnitude is positive and substantial for wins. This is consistent with the finding that voters underreact to wins, since it implies more accurate voters react more strongly to wins. It also provides evidence that voters do not improve their accuracy by minimizing weekly modifications to their rankings.

The common feature among almost all of these findings is indeed that saliency, or noticeability, is positively associated with overreaction. The relationships are sometimes subtle. For instance, home-field advantage is a non-salient characteristic of the signal. Thus, we would expect voters to be unresponsive to its significance—increasing the likelihood of win and decreasing that of loss—and thus (relatively) overreact to home wins and away losses, which is what we find. We might expect score margin to be more salient, and consequently exacerbate overreaction as it increases, yet we find this is true only for losses in general (SDEV is negative for wins). Overreaction does increase with SDEV,

however, for wins against ranked opponents. This story is still consistent with saliency then if we infer wins (and corresponding margins of victory) against non-ranked teams are non-salient, which is very plausible. While voters are not responsive to the scores of wins against weaker teams they should be; the best teams do in fact beat teams of all types by the greatest margins (show data?).

We can also attribute some of these findings to saliency of the priors, or really a lack thereof. Losses by top 10 teams are the most salient observable signal, and the TOP10 estimate for losses illustrates this. However, losses for all top 25 teams are highly noticeable, so one would struggle to argue that reactions to top 10 losses should be greater than those for teams ranked 11-25 due to saliency. The difference between the two rank groups is actually the priors. The data clearly show that the priors are much stronger for top 10 teams than other ranked teams, relative to those teams below them. For example, 91.4% of teams ranked in the top 10 in the preseason poll finished in the voter's top 25. The corresponding numbers for teams ranked 11-25, and not ranked at all but ranked by at least one other voter are 47.2% and 28.1%. This means that while top 10 teams were almost guaranteed to be very good (finish in the top 25), the difference between 11-25 and the best unranked teams was not so large. These differences are similar or stronger for later weeks. Other things equal this implies responses to losses by top 10 teams should be relatively smaller, since voters should have relatively greater confidence in the qualities of these teams. If voters did not make this (non-salient) distinction among relative qualities, however, they would react equally to losses by high and low ranked teams. Equal reactions to losses would appear as greater overreaction to losses by top 10 teams, which is exactly what we observe.

Voter-level heterogeneity is perhaps best analyzed graphically due to the limited sample. To do this I plot histograms of the fixed effects from regressions specified by equation (2), dropping the fixed effect-week interaction variables (this has little effect on the estimates for the other coefficients). I drop these terms so that we can more clearly examine average overreaction by voter. Figures 2 and 3 present these histograms for wins and losses, respectively. In the former, there is little evidence of interesting heterogeneity, as the variation in the estimates is symmetric and rather bell-shaped; essentially what we would expect to obtain from random variation in the sample (the average standard error of the estimates is 0.28). In the latter, the distribution appears skewed and the right tail is relatively fat (the average standard error of these estimates is 0.60). This is somewhat stronger evidence of real heterogeneity—that some voters overreact on average more than others, and that variation in these tendencies increases as the salience of the signals increases, since losses are more salient than wins.

## 4 Two Robustness Checks

In this section I discuss two robustness checks for the preceding analysis. The first is simple: I replace the estimated Bayesian updated ranks with the true ranks (the individual voters' final ranks).<sup>18</sup> The true ranks are unbiased estimators of the Bayesian ranks, but measured with substantial error. Thus we would expect our estimates of the effects of signal and prior characteristics on over or underreaction to be much less precise, but qualitatively similar, with this method. We also lose precision due to loss of sample, because the adjustment to account for teams leaving the poll is more severe in this case. The upside to using this approach is that it eschews reliance on any assumptions made to estimate the prior and signal distributions. With this approach I find strong evidence supporting the main results: that overreaction increases as the saliency of the signal and/or prior increases.

The second robustness check is methodologically much more complex, and qualitatively very different. Instead of explicitly computing Bayesian posterior ranks based on estimated prior beliefs, I *assume* the actual posteriors are Bayesian, and use these to estimate the voters' priors. I then compare the estimated priors to the data, and analyze overreaction again using the predicted posteriors computed with the new estimated priors. Once more I find evidence supporting our earlier conclusions.

### 4.1 Check 1: Replacing Estimated Posteriors with Truth

Implementing this check is very simple. I again use Definition 3.1 for the variable of interest, overreaction, only replacing  $r_{\Delta}^E$  with  $r_{\Delta}^F$ , in which  $r_{\Delta}^F$  is the (adjusted) prior rank minus the (adjusted) voter final rank. The main difficulty with doing this is that the adjustment for teams entering and exiting the polls is more costly now. Instead of 0-5 teams exiting the poll per voter per week, it is usually 5-10 teams.

Still, the results are strong. The mean overreactions for wins, losses and byes respectively are -1.52, 0.68 and -4.08. These effects are of the same sign, and larger in magnitude, than those found for the analysis discussed in section 4. They support the main findings that voters underreact to less salient signals (wins and byes) and overreact to more salient ones. I believe the magnitudes are larger here due to the heterogeneity of priors, and (relative) homogeneity of final ranks. Most voters' beliefs, about most teams, are fairly far from truth during the first half of the season, erring either on the side of pessimism or optimism. Their optimal Bayesian posteriors will adjust only slightly week to week, but their actual errors will be greater in magnitude when comparing to truth.

I also estimate equation (2) using the new measure of overreaction as the dependent variable. Results are presented in Table 6; they are similar to, but weaker than, those in Table 5. Saliency is still sometimes significant (HOME and TOP10 for losses), but not nearly as often (it is not for SDEV,

---

<sup>18</sup>This methodology is analogous to that used by Amir and Ganzach, as they compare analysts' forecasted earnings changes to actual changes.

APDEV etc.). TOP10 is negative and significant for wins, indicating voters underreact especially strongly to wins by top 10 teams. This is because actual ranks tend to remain unchanged after wins by these teams, while their final ranks are in expectation worse than current ranks due to the skewed nature of the data (top 10 teams are much more likely to fall than rise in rank). One other interesting difference is that the home bias variables (ST and REG) are no longer significant in either equation. I believe these differences are due to these estimates being less precise because of measurement error. Overall, though, they support the findings of Section 3.

## 4.2 Check 2: Estimating the Bayesian Priors

As a final robustness check for the individual voter analysis I invert my approach: rather than estimate the Bayesian posteriors and compare these to the actuals, I assume the actuals are in fact Bayesian, and estimate the priors that are most consistent with (*rationalize*) the data. Then, I first compare the estimated priors to the empirical frequencies of final rank conditional on current (prior) rank. Second, I compare the predicted Bayesian posteriors, obtained using the newly estimated priors, to the actuals. Doing this I find the estimated priors are inconsistent with both the data and weak theory—they do not become ‘stronger’ as the season progresses as we would expect given the acquisition of information. In addition, I find that overreaction, defined using predicted Bayesian posteriors, is still positively related to salience (losses, home field, etc.).

### 4.2.1 Method

To estimate the priors that rationalize the observed behavior I continue to use the model specified in equation (1), and the same score distributions, rank groups, and method for accounting for limited rankings and mapping beliefs to rankings as used above. I am comfortable continuing to use these assumptions since they are all fairly weak, and the first robustness check abstracted from many of them. Instead of using the estimated priors from section 3, however, I do a grid search over a large set of priors to find those that minimize the distance between predicted Bayesian posterior rankings and actuals. I search for a different set of priors for each week, so seven sets of priors total, since we expect the priors to change week to week. Formally, I search for

$$f_t^* = \underset{f_t}{\operatorname{argmin}} \sum_i \sum_j |r_{ijt}(\hat{f}_t) - \tilde{r}_{ijt}|, \quad t = 1, \dots, 7, \quad (6)$$

in which  $\hat{r}_{ijt}(f_t)$  denotes the predicted rank of team  $i$  for voter  $j$  in week  $t$  as a function of the priors,  $f_t$ , and other observables, and  $\tilde{r}$  is actual posterior. I use sum of absolute deviations as a distance metric to minimize sensitivity to outliers.

The key step here is the choice of priors to search over. The set of possibilities is infinite. It is also

made more complicated as the term ‘priors’ as I have used it actually refers to a joint distribution of order statistics. Specifically, we require the probability an order statistic falls in one of five categories (the four rank groups comprising the top 25 and unranked, or 26-plus) for teams of each prior rank group, of which there are 32 (the top 25 and the seven categories of unranked teams described in the section on prior estimation above). Thus each joint prior distribution is a  $32 \times 5$  matrix, with rows that all sum to one and columns which, when weighted by the appropriate number of teams, sum to the number of teams in each rank group.

To generate these joint distributions of order statistics I assume the underlying variable on which teams are ranked is jointly normal, with zero covariance across teams. For concreteness I refer to this variable here as quality. I assume there are 119 teams (the number of Division I-A teams in 2005) and equally space the means of the quality variable across the unit interval  $(0, 0.0085, \dots, 0.9915, 1)$ . I again divide the teams into rank categories. I use the same categories for the top 25 teams as used above (1-6, 7-12, ...), and divide the unranked teams into two groups, ‘others receiving votes’ (ranks 26-37) and all others (ranks 38+). I create the others receiving votes category because it is a distinct group of teams—those just on the cusp of being in the top 25 (there are an approximate average of 12 teams receiving votes throughout the first half of the 2006 season). There are thus six rank categories.

I then assume that variance is constant for all teams in a rank category, and permute five different variances (0.002, 0.006, 0.010, 0.014, 0.018) over the six rank groups. Thus I generate a set of  $5^6 = 15,625$  joint prior distributions to search over, each of which I estimate through simulation (10,000 runs). This method of determining the prior set that I search over is fairly arbitrary. It serves our purposes, however, as the means and variances of the quality variable are not parameters of interest. What is critical is that we have a large and varied set of order statistics to search over, which use of this method yields.

#### 4.2.2 Results

Tables 7 and 8 present some properties of the estimated Bayesian priors in comparison with the analogous empirical frequencies. Table 7 presents expected final rank for each of the four top 25 prior rank groups, given that the final rank is in the top 25. I condition on final rank being in the top 25 to avoid having to make an assumption about the expected rank of unranked teams. We would expect to see these expected ranks decreasing as the season progresses and information is obtained about the teams, especially for top-ranked teams. That is, we would expect a greater fraction of teams ranked 1-6 in week 7 to also be 1-6 in the final poll than the fraction of teams 1-6 in week 1 who finish 1-6. We do observe this trend in the empirical frequencies—the expected rank of teams with prior rank 1-6 decreases from 9.55 in week 1 to 7.14 in week 7. However, using the estimated priors these ranks actually increase, from 6.62 to 8.87. These estimates are likely due to the fact that some top ranked

teams lost in week 7, while they did not in week 1. If voters overreact to losses and underreact to wins, assuming they are Bayesian will lead us to estimate stronger priors when teams win more frequently (as in week 1), and weaker priors when teams lose (as in week 7).

The numbers in Table 8 are even more striking. These are simply the probabilities of finishing unranked, which we certainly expect to decrease for all teams in the top 25 as the season progresses. We see this phenomenon fairly clearly in the actuals, as the percentages decrease from week 1 to 7 for all teams except those ranked 13-18. The pattern is almost the opposite for the estimates, as the percentages increase for 19 out of 25 prior ranks. This is strongly suggestive of non-Bayesian behavior, consistent with underreaction to wins, since most top 25 teams won in week 1.

Finally, I again conduct the overreaction regression analysis discussed above. I continue to use Definition 3.1 to define the dependent variable, overreaction, now using the predicted posterior ranks as the estimated Bayesian posteriors. Overreaction is essentially now a reaction residual. If the voters were in fact Bayesian, this residual would be independent of such observable variables as win, loss, etc. Instead we find a pattern similar to that found above in which overreaction is significantly greater for losses: the mean of redefined OVER is 0.30, 0.99, and 0.27 for wins, losses and byes respectively. Even though we have fit the priors to the data we find evidence of overreaction to more salient signals. The values are all positive because the predicted rank changes regress to the mean (of 0) due to the nature of the estimator. I also, again, estimate equation (2) for wins and losses separately, using the newly defined OVER as the dependent variable. I do not report the results, but find several variables to be significant, which should have coefficients of zero under rationality, including HOME for both wins and losses. This again supports the results found above.

## 5 The Aggregate Polls

The analysis thus far has focused on the behavior of individual voters over a relatively short time period, the 2006 season, due to data limitations. The data on aggregate voter behavior are publicly available for all weeks and years. These data, while not as valuable for analyzing individual behavior, are still worth exploring for various reasons. First, we can estimate priors using historical final rank frequencies. Second, we can conduct the analyses for other years, and confirm that any trends found are not confined to 2006 for some reason. Third, we observe rankings for a larger set of teams—30-50 for most weeks, since we observe aggregate point totals for all teams receiving votes, including those not in the top 25. This allows for a wider range of reactions, especially for teams ranked 11-25, i.e. these teams now have farther to potentially fall.

I examine weeks 1-4 of the 2004-2006 seasons. I include the 2006 season even though we have already analyzed it so that we can see how the aggregate approach changes results for a particular

season. I restrict the sample to weeks 1-4 because, as mentioned above, using data further from the final poll increases power and minimizes the probability of voters committing the hot hand fallacy (inferring trend in team quality change). I use the same score distributions and rank categories for the top 25 as above, but now add an additional final rank category, 26-35. I limit the maximum final rank to 35 even though it varies year to year because there are at least 35 teams in the final poll in every year since the poll began including 25 teams (1989). I estimate the prior distributions using the historical final rank frequencies for seasons 1989 through the year prior to current season (2003 for the 2004 estimates, 04 for 05 etc.). These frequencies are still quite noisy and require smoothing to obtain monotonically increasing expected rank. I use the same criterion and method for smoothing (smooth until minimum 13 consecutive increasing expected rank, then average). I discuss an alternative where I do not smooth at all below.

The statistic used previously for testing validity of the rankings, mean absolute deviation from final rank, is somewhat less impressive in this case: 6.3, 6.33 and 6.46 for the actuals, estimates and priors respectively. While both actual and estimated posteriors are superior to the priors, the margin is small, and the estimates no longer out-perform the actuals. Formal tests are easier now as the sample is now much closer to i.i.d., since there are no repeated games or voters. Paired t-tests for two-sided tests are insignificant at conventional levels. I view this as evidence that the aggregate rankings are superior to the individual ones, which is not surprising. The estimated rankings are actually less accurate in 2006 than 05. This is reassuring, because it implies the validity of the individual analysis demonstrated above is not dependent on anomalies that occurred during the 2006 season.

Overreaction, measured in the usual way, now has means of -0.60, -0.07 and -1.00 for wins, losses and byes respectively. While the overall pattern of relatively stronger overreaction to losses continues to hold, it appears the average voter no longer ever overreacts. This is misleading, however, as we see when we break out the statistic by initial rank group. Mean overreaction to losses for top 10, 11-25 and 26+ teams are 3.06, -2.15 and 0.29, respectively. Overreaction is still strong on average to losses for top teams. But it is actually very low, i.e. there is strong underreaction, for middle ranked teams because the priors for these teams are so uncertain, and now the potential rank changes are greater. That is, a team that was ranked 20 could only have a minimum rank change of -5 before, but now its minimum is -10 or lower. Still, it is interesting that on average voters appear to be Bayesian in their reactions to losses. This suggests that while people may make particular mistakes in various circumstances, the mistakes may cancel out and thus be rationalizable in a sense.

Regression analysis yields similar results. I again estimate equation (2), now dropping voter-specific variables, adding year fixed effects and a new dummy, TOP11\_25, equal to 1 if prior rank is 11-25. Results are presented in Table 9. For wins, voters, even in the aggregate, underreact to lack of home-field advantage, and score margin. They overreact to score margin when it is salient, i.e. the opponent

is ranked in the top 25. Regarding losses, the most important variables are TOP10 and TOP11.25, which are associated with prior strength and weakness, respectively. Voters overreact to losses when the priors are strong and do the opposite when weak. Again, this can be interpreted as related to salience, as the variation in prior strength is not salient and consequently not appreciated by the voters.

As a robustness check for these results, I also estimate the posteriors without smoothing the priors. I simply use the historical final rank frequencies as priors, even though as noted these are very noisy. Doing this the estimated posterior mean absolute deviation from truth is 6.49, which is actually greater than the mean deviation for the priors (still 6.46), though not significantly so. Still, the overreaction measures are qualitatively similar to those found elsewhere. The mean overreactions are -0.76, 0.20, and -2.37 for wins, losses and byes respectively. For losses, it is 2.81, -2.33, and 0.96 for teams in the top 10, 11-25 and 26+ respectively. In the regression analysis, due to the imprecision of the estimated posteriors we no longer obtain significant estimates for HOME or SDEV, but still obtain a significant negative estimate for TOP11.25 in the losses model.

## 6 Concluding Remarks

This study has compiled extensive evidence of systematic over and underreaction to new information by experts in a real-world context. The errors are not large in an absolute sense, but are statistically significant. Factors that appear to make overreaction more likely, regarding either the prior or signal, are characterized by salience, or visibility. Namely, voters improve teams' rankings excessively after wins at home and over ranked opponents by large margins. They worsen rankings excessively after losses on the road, by large margins, and in general for top ranked teams. Wins against unranked teams, and the margins of those wins, are unappreciated by voters. The reasons why salience affects the likelihood of overreaction are not obvious, and are outside the scope of this paper.

## References

- AMIR, E., AND Y. GANZACH (1998): "Overreaction and underreaction in analysts forecasts," *Journal of Economic Behavior and Organization*, 37(3), 333–347.
- BARBERIS, N., AND R. THALER (2002): "A Survey of Behavioral Finance," .
- CAMERER, C. (1989): "Does the Basketball Market Believe in the Hot Hand, '?,", *The American Economic Review*, 79(5), 1257–1261.
- CHARNESS, G., E. KARNI, AND D. LEVIN (2007): "Individual and Group Decision Making Under Risk:

- An Experimental Study of Bayesian Updating and Violations of First-order Stochastic Dominance,” *Journal of Risk and Uncertainty*, 35(2), 129–148.
- DEBONDT, W., AND R. THALER (1985): “Does the stock market overreact,” *Journal of Finance*, 40(3), 793–805.
- DELLAVIGNA, S. (2007): “Psychology and Economics: Evidence from the Field,” Working Paper.
- GOFF, B. (1996): “An assessment of path dependence in collective decisions: evidence from football polls,” *Applied Economics*, 28(3), 291–297.
- GRETHER, D. (1980): “Bayes Rule as a Descriptive Model: The Representativeness Heuristic,” *The Quarterly Journal of Economics*, 95(3), 537–557.
- HONG, H., AND J. STEIN (1999): “A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets,” *The Journal of Finance*, 54(6), 2143–2184.
- LEBOVIC, J., AND L. SIGELMAN (2001): “The forecasting accuracy and determinants of football rankings,” *International Journal of Forecasting*, 17(1), 105–120.
- LEVITT, S., AND J. LIST (2007): “What Do Laboratory Experiments Tell Us About the Real World?,” *Journal of Economic Perspectives*, forthcoming.
- LOGAN, T. (2007): “Whoa, Nellie! Empirical Tests of College Football’s Conventional Wisdom,” .
- OFFERMAN, T., AND J. SONNEMANS (2004): “What’s Causing Overreaction? An Experimental Investigation of Recency and the Hot-hand Effect,” *Scandinavian Journal of Economics*, 106(3), 533–554.
- RABIN, M. (1998): “Psychology and Economics,” *Journal of Economic Literature*, 36(1), 11–46.
- TVERSKY, A., AND D. KAHNEMAN (1974): “Judgment under Uncertainty: Heuristics and Biases,” *Science*, 185(4157), 1124.

## A Estimation

### A.1 Quality Changes

To illustrate why performance is correlated with time if: 1) voters rank teams on current quality and, 2) quality changes throughout the season, consider the following simple example. Suppose there are only two teams, two games, and performance is a deterministic function of quality. Specifically, suppose team  $i$ 's quality in period  $t$  is  $X_{i,t}$  and the score between teams 1 and 2 in period  $t$  is  $s_{12}^t = X_{1,t} - X_{2,t}$ . Suppose we have a sample of data from many seasons, and  $X_{1,1}$  and  $X_{2,1}$  are i.i.d. for all seasons. So that this example is analogous to the results presented in Table 1 suppose team 1 is always the home team. Last, suppose team qualities may change so that  $X_{i,2} = X_{i,1} + \epsilon_i$ , in which  $\epsilon_i$  is i.i.d. with mean 0 and positive variance.

Let  $\tilde{r}_i$  denote team  $i$ 's final rank. Voters rank teams on current quality so  $\tilde{r}_1 = 1 \leftrightarrow s^2 > 0$  (w.l.o.g. ignore ties). Then the expected score margin of the second and final game given that the home team is ranked 1 and the away team ranked 2 is  $E(s^2 | s^2 > 0) = E(X_{i,1} + \epsilon_i - X_{j,1} - \epsilon_j | X_{i,1} + \epsilon_i - X_{j,1} - \epsilon_j > 0)$ . The expected score margin of the first game given these final ranks is  $E(s^1 | s^2 > 0) = E(X_{i,1} - X_{j,1} | X_{i,1} + \epsilon_i - X_{j,1} - \epsilon_j > 0)$ . Since  $E(\epsilon_i | X_{i,1} + \epsilon_i - X_{j,1} - \epsilon_j > 0) > 0$  and  $E(\epsilon_j | X_{i,1} + \epsilon_i - X_{j,1} - \epsilon_j > 0) < 0$  we see that  $E(s^1 | s^2 > 0) < E(s^2 | s^2 > 0)$ . This implies that observed performance is correlated with time. The logic of this example applies when there are greater than two teams and games.

### A.2 Score Distributions

Different score margin distributions are estimated for home-away and away-home games of each (true) rank group. There are seven true rank groups (1-6, 7-12, 13-18, 19-25, ranked in previous two years, ranked in previous three-five years, unranked in previous five years). There are 17 points of support for each score margin distribution (-50-, [-49,-43],...,[-7,-1],0,[1,7],..., [43,49],50+). Let  $c_{j,k}^i$  denote the historical count of games with score margins in category  $i \in \{1, \dots, 17\}$  for games between home team of rank group  $j$  and away team of rank  $k$ .  $j$  and  $k$  are henceforth suppressed. For  $i \in \{3, \dots, 6, 12, \dots, 15\}$  let  $\tilde{c}^i = \frac{1}{5} \sum_{\hat{i}=i-2}^{i+2} c^{\hat{i}}$ . For  $i \in \{1, 2\}$  let  $\tilde{c}^i = \frac{1}{3+I(i=2)} \sum_{\hat{i}=1}^{i+2} c^{\hat{i}}$ , in which  $I(i=2) = 1$  if  $i = 2$ , else  $I(i=2) = 0$ . For  $i \in \{16, 17\}$  let  $\tilde{c}^i = \frac{1}{3+I(i=16)} \sum_{\hat{i}=i-2}^{18} c^{\hat{i}}$ .  $\tilde{c}^i = c^i$  if  $i = 0$ . Let  $g(s_{jk}^i)$  denote the probability the score margin,  $s$ , is in category  $i$  for games between home teams in rank group  $j$  and away in group  $k$ . Then  $\hat{g}(s_{jk}^i) = \frac{\tilde{c}^i}{\sum_{\hat{i}} \tilde{c}^{\hat{i}}}$ .

### A.3 Prior Distributions

There are again seven final rank groups and there are 32 prior rank groups, which include the 25 ranked teams and 7 groups of unranked teams.<sup>19</sup> The priors for the unranked teams are estimated as the raw empirical frequencies of final rank. Let  $n_{j,t,k}^v$  denote the number of teams for voter  $v$  in prior rank group  $j$  and period  $t$  that finished in rank group  $k$  (for that voter). Then if  $j > 25$  (the team is currently unranked),  $f_{i,t}(k|r_{it} \in j)$  is estimated as  $\frac{\sum_v n_{j,t,k}^v}{\sum_k \sum_v n_{j,t,k}^v}$ .

If  $i \leq 25$  I smooth the estimated distributions using neighboring distributions. That is, I average the frequencies for all prior rank with similar prior ranks (other prior ranks just above and below). Specifically, letting  $b$  denote the number of neighbors, I estimate  $f_{i,t}(k|r_{it} = j)$  as  $\frac{\sum_{\hat{j}=j-b:j+b} \sum_v n_{\hat{j},t,k}^v}{\sum_{\hat{j}=j-b:j+b} \sum_k \sum_v n_{\hat{j},t,k}^v}$ . I ‘wrap around’ neighbors for neighbors below  $j = 1$  and above  $j = 25$ . That is, I replace  $j - b = 0$  with 1,  $j - b = -1$  with 2, etc., and  $j + b = 26$  with 25,  $j + b = 27$  with 24, etc. This satisfies ‘rank accounting’ constraints.

I use the minimum  $b$  such that expected rank, calculated using the estimated priors, monotonically increases for at least 13 consecutive prior ranks ( $j$ ’s). Denote this minimum  $b$  as  $\hat{b}$ . The number 13 is arbitrary; it is chosen because it is the mid-point of the top 25. Ideally all 25 expected ranks would monotonically increase, however, the minimum number of neighbors needed to satisfy this requirement causes extreme over-smoothing for some prior ranks.

If there is  $j$  such that estimated  $E_t(r|r_{it} = j) > E_t(r|r_{it} = j + 1)$ , i.e. monotonicity is violated after smoothing, I simply average the estimated priors for  $j$  and  $j + 1$  (and  $j + 2$ , etc., as necessary):

$$\hat{f}_{i,t}(k|r_{it} = j) = \hat{f}_{i,t}(k|r_{it} = j + 1) = \frac{\sum_{j,j+1} \left[ \sum_{\hat{j}=j-b:j+b} \sum_v n_{\hat{j},t,k}^v \right]}{\sum_{j,j+1} \left[ \sum_{\hat{j}=j-b:j+b} \sum_k \sum_v n_{\hat{j},t,k}^v \right]}. \quad (7)$$

This is an admittedly crude way to obtain weakly monotone increasing estimated expected ranks. However, it accurately reflects the fact that the voters should be very uncertain about the true ranks of teams of some prior ranks, given the extreme variability of actual final ranks of those teams. In addition, we no longer precisely satisfy the rank accounting constraints. The errors, however, are small, and I believe that the benefit of the simplicity of this method outweighs the cost of imprecision.

---

<sup>19</sup>See ‘Limited Rankings’ subsection.

## B Tables

Table 1: p-values for two-sided t-tests of  $H_0$ : expected score differences conditional on final rank are constant throughout the season

| HomeRk | AwayRk   | Period        | n   | $\bar{s}$ | $\hat{\sigma}_{\bar{s}}$ | p-value |
|--------|----------|---------------|-----|-----------|--------------------------|---------|
| 1-12   | 13-25    | Aug-Oct 15    | 69  | 13.8      | 1.7                      | 0.17    |
| 1-12   | 13-25    | Oct 16-Dec 15 | 86  | 17.1      | 1.7                      |         |
| 1-12   | Unranked | Aug-Oct 15    | 140 | 21.3      | 1.4                      | 0.79    |
| 1-12   | Unranked | Oct 16-Dec 15 | 150 | 20.7      | 1.3                      |         |
| 13-25  | 1-12     | Aug-Oct 15    | 70  | -6.9      | 1.8                      | 0.18    |
| 13-25  | 1-12     | Oct 16-Dec 15 | 65  | -3.5      | 1.7                      |         |
| 13-25  | Unranked | Aug-Oct 15    | 161 | 14.9      | 1.2                      | 0.28    |
| 13-25  | Unranked | Oct 16-Dec 15 | 173 | 13.0      | 1.2                      |         |

Notes: 'Rk' = final AP aggregate rank;  $\bar{s}$  = mean home score - away score;  $\hat{\sigma}_{\bar{s}}$  = estimated std error of  $\bar{s}$ . Sample includes games played 1989-2005 with at least one Division I-A team on non-neutral field. 'Unranked' restricted to teams receiving votes in final aggregate poll in at least one of previous two seasons.

Table 2: Individual Voter Final Rank Frequencies Conditional on Preseason Rank (2006)

| Prior | Pr( $r \in [1,6]$ ) | Pr( $r \in [7,12]$ ) | Pr( $r \in [13,18]$ ) | Pr( $r \in [19,25]$ ) | Pr( $r \in [26+]$ ) | Total | E[r] |
|-------|---------------------|----------------------|-----------------------|-----------------------|---------------------|-------|------|
| 1     | 53.1%               | 23.4%                | 18.8%                 | 4.7%                  | 0.0%                | 100%  | 8.0  |
| 2     | 29.7%               | 26.6%                | 28.1%                 | 14.1%                 | 1.6%                | 100%  | 11.6 |
| 3     | 34.4%               | 32.8%                | 23.4%                 | 7.8%                  | 1.6%                | 100%  | 10.2 |
| 4     | 18.8%               | 46.9%                | 21.9%                 | 7.8%                  | 4.7%                | 100%  | 11.9 |
| 5     | 32.8%               | 39.1%                | 21.9%                 | 4.7%                  | 1.6%                | 100%  | 9.8  |
| 6     | 42.2%               | 34.4%                | 14.1%                 | 3.1%                  | 6.3%                | 100%  | 9.8  |
| 7     | 50.0%               | 28.1%                | 9.4%                  | 6.3%                  | 6.3%                | 100%  | 9.4  |
| 8     | 48.4%               | 23.4%                | 15.6%                 | 4.7%                  | 7.8%                | 100%  | 10.1 |
| 9     | 18.8%               | 31.3%                | 25.0%                 | 6.3%                  | 18.8%               | 100%  | 15.4 |
| 10    | 23.4%               | 20.3%                | 15.6%                 | 3.1%                  | 37.5%               | 100%  | 19.0 |

Notes: Prior = preseason rank. E(r) computed using midpoint ranks from each final rank category ([1,6],...) and 35 for unranked category.

Table 3: Absolute Deviations from Postseason Ranks

|      | Estimates | Actuals | Priors | FlatPrior |
|------|-----------|---------|--------|-----------|
| mean | 3.39      | 3.61    | 3.73   | 5.09      |
| sd   | 3.19      | 3.32    | 3.43   | 2.96      |
| N    | 10893     | 10893   | 10893  | 10893     |

Table 4: Estimated and Actual Rank Changes (Prior - Posterior Rank)

|      | <b>Wins: Est.</b> | <b>Actual</b> | <b>Losses: Est.</b> | <b>Actual</b> | <b>Byes: Est.</b> | <b>Actual</b> |
|------|-------------------|---------------|---------------------|---------------|-------------------|---------------|
| mean | 1.31              | 1.15          | -4.32               | -4.76         | -1.91             | 0.43          |
| sd   | 3.15              | 2.29          | 3.51                | 3.63          | 3.18              | 1.49          |
| N    | 8028              | 8028          | 2008                | 2008          | 857               | 857           |

Table 5: Estimation results: Wins and Losses (Std Errors Clustered by Game)

| <b>Wins</b> | <b>Coefficient</b> | <b>(Std. Err.)</b> | <b>Losses</b> | <b>Coefficient</b> | <b>(Std. Err.)</b> |
|-------------|--------------------|--------------------|---------------|--------------------|--------------------|
| HOME        | 1.449 **           | (0.315)            | HOME          | -1.267 *           | (0.587)            |
| WEEK        | -0.066             | (0.122)            | WEEK          | -0.019             | (0.277)            |
| SDEV        | -0.048 **          | (0.009)            | SDEV          | -0.047 †           | (0.027)            |
| TOP10       | -0.945 **          | (0.310)            | TOP10         | 4.779 **           | (0.654)            |
| OPP25       | 0.059 **           | (0.017)            | OPP25         | 0.018              | (0.020)            |
| SDEVOPP25   | 0.100 **           | (0.037)            | SDEVOPP25     | 0.035              | (0.027)            |
| APDEV       | -0.069 **          | (0.026)            | APDEV         | 0.108 **           | (0.038)            |
| RKSD        | -0.515 **          | (0.124)            | RKSD          | 0.192              | (0.202)            |
| ST          | 0.233 †            | (0.125)            | ST            | 0.180              | (0.296)            |
| REG         | 0.214 *            | (0.096)            | REG           | 0.137              | (0.226)            |
| YRS         | -0.202 *           | (0.095)            | YRS           | 0.259              | (0.284)            |
| ACCURACY    | 0.608              | (0.456)            | ACCURACY      | -0.012             | (1.196)            |

Significance levels : † : 10% \* : 5% \*\* : 1%

Table 6: Estimation results : Robustness Check 1 (Std Errors Clustered by Game)

| <b>Wins</b> | <b>Coefficient</b> | <b>(Std. Err.)</b> | <b>Losses</b> | <b>Coefficient</b> | <b>(Std. Err.)</b> |
|-------------|--------------------|--------------------|---------------|--------------------|--------------------|
| HOME        | -0.078             | (0.647)            | HOME          | -2.221 **          | (0.780)            |
| WEEK        | 0.047              | (0.217)            | WEEK          | -0.108             | (0.216)            |
| SDEV        | 0.001              | (0.020)            | SDEV          | 0.044              | (0.028)            |
| TOP10       | -1.731 *           | (0.665)            | TOP10         | 2.354 *            | (0.911)            |
| OPP25       | -0.068 *           | (0.035)            | OPP25         | 0.022              | (0.029)            |
| SDEVOPP25   | 0.026              | (0.056)            | SDEVOPP25     | 0.011              | (0.026)            |
| APDEV       | 0.015              | (0.052)            | APDEV         | 0.032              | (0.069)            |
| RKSD        | -0.521 *           | (0.223)            | RKSD          | -0.675 *           | (0.267)            |
| ST          | 0.354              | (0.264)            | ST            | -0.136             | (0.339)            |
| REG         | -0.033             | (0.186)            | REG           | 0.024              | (0.338)            |
| YRS         | 0.120              | (0.180)            | YRS           | -0.152             | (0.356)            |
| ACC         | -0.198             | (0.679)            | ACC           | -1.602             | (1.855)            |

Significance levels : † : 10% \* : 5% \*\* : 1%

Table 7: E(Final Rank|Final Rank in Top 25)

|        | Prior Rk | Week 1       | Week 2 | Week 3 | Week 4 | Week 5 | Week 6 | Week 7       |
|--------|----------|--------------|--------|--------|--------|--------|--------|--------------|
| Est.   | 1-6      | <b>6.62</b>  | 6.71   | 8.87   | 6.31   | 6.92   | 8.68   | <b>8.27</b>  |
|        | 7-12     | <b>10.45</b> | 11.60  | 11.05  | 10.89  | 10.41  | 10.35  | <b>12.17</b> |
|        | 13-18    | <b>15.73</b> | 14.74  | 12.52  | 14.41  | 14.34  | 13.50  | <b>12.69</b> |
|        | 19-25    | <b>16.12</b> | 16.11  | 17.06  | 17.32  | 17.15  | 16.99  | <b>17.24</b> |
| Actual | 1-6      | <b>9.55</b>  | 9.21   | 8.22   | 6.78   | 6.70   | 6.56   | <b>7.14</b>  |
|        | 7-12     | <b>9.18</b>  | 9.58   | 10.88  | 12.41  | 12.46  | 11.92  | <b>13.71</b> |
|        | 13-18    | <b>14.23</b> | 15.70  | 15.53  | 17.15  | 16.54  | 15.30  | <b>11.66</b> |
|        | 19-25    | <b>18.24</b> | 15.79  | 16.19  | 15.63  | 13.33  | 14.07  | <b>12.96</b> |

Notes: Est. = priors estimated assuming voters use Bayes' rule; Actual = 2006 empirical frequencies of final rank given prior (current, in week  $t$ ) rank

Table 8: Pr(Final Rank = 26+ )

|        | Prior Rk | Week 1       | Week 2 | Week 3 | Week 4 | Week 5 | Week 6 | Week 7       |
|--------|----------|--------------|--------|--------|--------|--------|--------|--------------|
| Est.   | 1-6      | <b>0.0%</b>  | 0.0%   | 3.2%   | 0.0%   | 0.0%   | 4.2%   | <b>1.9%</b>  |
|        | 7-12     | <b>9.8%</b>  | 0.8%   | 11.9%  | 0.2%   | 11.4%  | 17.7%  | <b>1.2%</b>  |
|        | 13-18    | <b>5.3%</b>  | 18.7%  | 25.4%  | 12.1%  | 12.0%  | 23.0%  | <b>29.3%</b> |
|        | 19-25    | <b>38.4%</b> | 43.5%  | 34.6%  | 32.2%  | 32.0%  | 41.5%  | <b>46.2%</b> |
| Actual | 1-6      | <b>2.6%</b>  | 3.4%   | 1.3%   | 0.3%   | 0.3%   | 0.0%   | <b>0.5%</b>  |
|        | 7-12     | <b>26.0%</b> | 29.4%  | 24.2%  | 17.7%  | 18.2%  | 20.8%  | <b>21.2%</b> |
|        | 13-18    | <b>44.0%</b> | 44.3%  | 48.7%  | 47.1%  | 44.0%  | 51.8%  | <b>50.0%</b> |
|        | 19-25    | <b>63.2%</b> | 60.5%  | 54.0%  | 50.9%  | 53.1%  | 48.4%  | <b>35.8%</b> |

Notes: Est. = priors estimated assuming voters use Bayes' rule; Actual = 2006 empirical frequencies of final rank given prior (current, in week  $t$ ) rank

Table 9: Estimation results : Aggregate Polls, 2004-2006 (Robust std errors)

| Wins        | Coefficient | (Std. Err.) | Losses  | Coefficient | (Std. Err.) |         |         |
|-------------|-------------|-------------|---------|-------------|-------------|---------|---------|
| HOME        | 0.934       | *           | (0.383) | HOME        | -1.130      | (0.689) |         |
| WEEK        | 0.169       |             | (0.163) | WEEK        | 0.820       | *       | (0.338) |
| SDEV        | -0.060      | **          | (0.012) | SDEV        | 0.080       | †       | (0.043) |
| OPP25       | 0.051       |             | (0.557) | OPP25       | -0.618      |         | (1.078) |
| SDEVOPP25   | 0.196       | **          | (0.075) | SDEVOPP25   | -0.132      | *       | (0.052) |
| TOP10       | 0.585       |             | (0.395) | TOP10       | 2.811       | **      | (0.911) |
| TOP11_25    | 0.322       |             | (0.444) | TOP11_25    | -3.072      | **      | (0.905) |
| _Iyear_2005 | 0.756       | †           | (0.418) | _Iyear_2005 | 0.553       |         | (0.811) |
| _Iyear_2006 | 0.398       |             | (0.396) | _Iyear_2006 | -0.623      |         | (0.797) |
| Intercept   | -2.091      | **          | (0.666) | Intercept   | -0.368      |         | (1.058) |

Significance levels : † : 10% \* : 5% \*\* : 1%

# C Figures

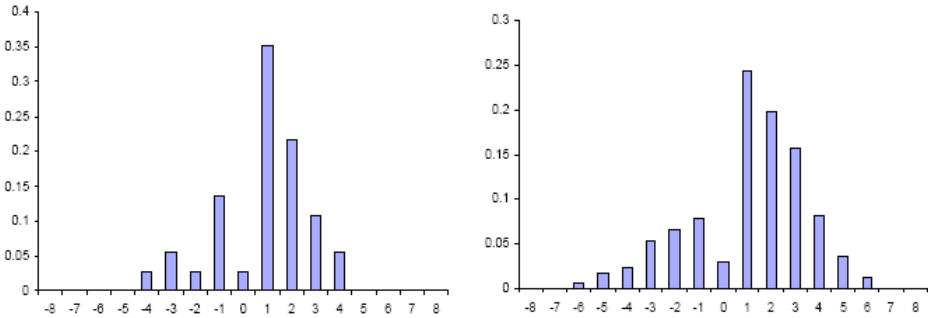


Figure 1: Unsmoothed (left), smoothed (right) score (home - away) histograms for games between teams with final rank 1-6

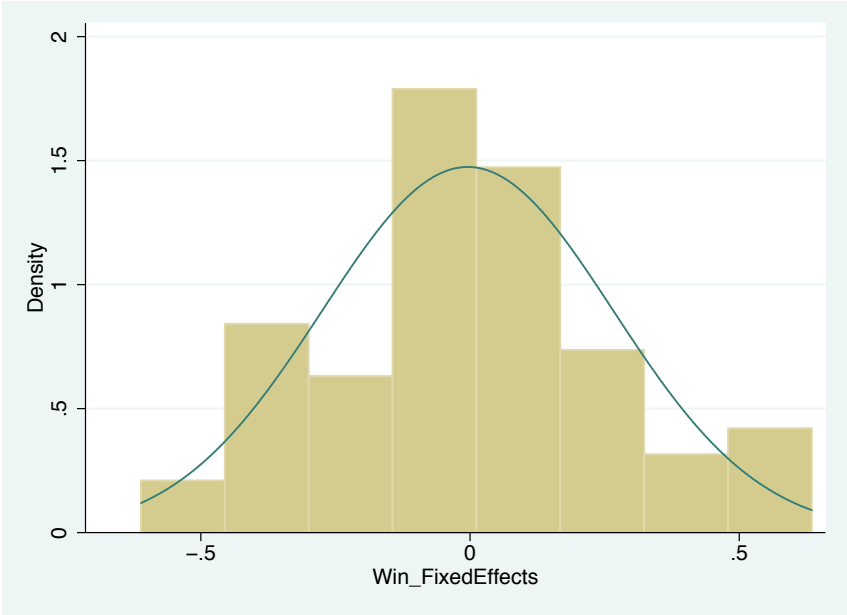


Figure 2: Wins: Histogram of Voter Fixed Effects with Normal PDF

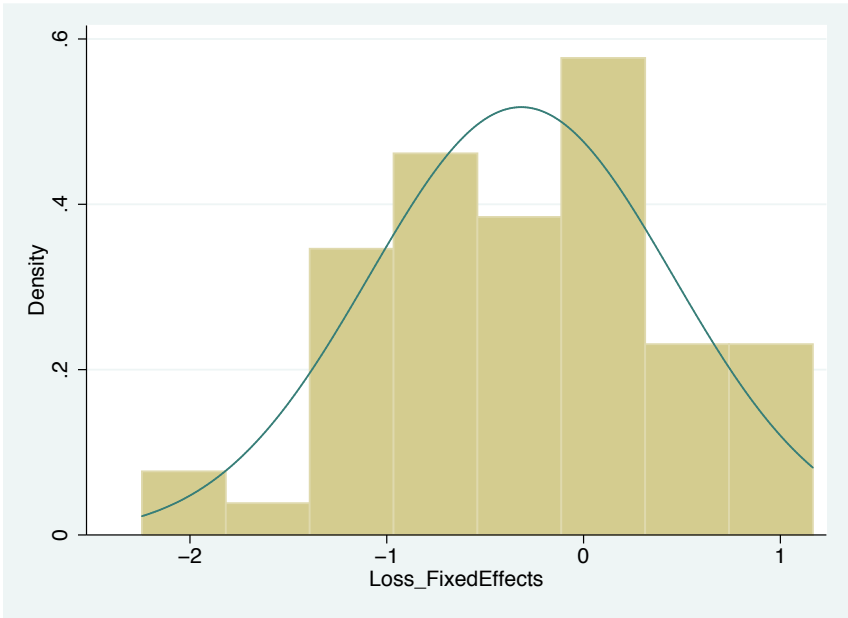


Figure 3: Losses: Histogram of Voter Fixed Effects with Normal PDF