Testing Bayesian Updating with the AP Top 25

Daniel F. Stone[*]

Johns Hopkins University

July 2008

Abstract

Most studies of Bayesian updating use experimental data. I use a novel, real-world data source–the Associated Press (AP) college football poll, a weekly subjective ranking of the top 25 teams–to test the validity of Bayes' rule as a descriptive model. I argue that the poll voters' individual final rankings represent their 'true' rankings, for a given season, and use historical score data and final rank frequencies to estimate benchmark Bayesian posterior rankings. I compare estimated Bayesian to observed ranking changes and find evidence of both Bayesian updating and systematic over and underreaction to new information. Overreaction is positively associated with the salience of new information, and lack of salience of strong priors. The evidence that voters overreact to losses by highly ranked teams is especially strong. The finding that salience drives both over and underreaction allows well known heuristics that appear to be in conflict, like representativeness and anchoring, to be reconciled. The results are confirmed using multiple methods and the aggregate polls.

JEL Classification Numbers: D80, D83, D84

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

1 Introduction

Most studies of Bayesian updating use experimental data.[1] While this research has led to tremendous insight into human behavior, it is inherently subject to a variety of criticisms. A common one is that in experimental settings agents lack expertise and the ability to learn. Many experiments attempt to address this concern by 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[2] and agents not having sufficient time to make optimal decisions.[3]

A perhaps more subtle weakness of experimental research is that it may be relatively unlikely to yield findings explaining a range of behavior. This is because experiments are usually set up to test hypotheses, formed prior to the start of the experiment, regarding particular behavioral anomalies. Hypotheses of non-Bayesian belief updating usually fall in either the category of overreaction or underreaction to new information. Overreaction occurs when individuals adjust their beliefs excessively in response to a signal, acting as if they put too much weight on the new information and/or insufficient weight on the prior. When found in experiments it has been ascribed to the *representativeness* and *availability* heuristics, *alarmism* and the *base-rate* fallacy. People underreact when they make the opposite mistake and revise their beliefs an insufficient amount; this has been attributed to, e.g., *over-confidence*, the *anchoring* and *confirmatory* biases.[4] Because most experimental studies concentrate on showing that one or the other bias exists, they are usually incapable of answering the question of which bias occurs when, and how the biases may be reconciled.

This paper contributes to this research area by using a novel, real-world data source to test Bayesian

---

[1]See, e.g., Tversky and Kahneman (1974), Grether (1980), Rabin (1998) and Delavande (2008). DellaVigna (2007) reviews field evidence of behavioral anomalies, and does not mention any studies that focus explicitly on belief updating.

[2]Charness, Karni, and Levin (2007) find strong evidence of the significance of this factor.

[3]Levitt and List (2007) provide an interesting discussion of the generalizability of experimental results given some of these issues. Viscusi (1985), for example, discusses evidence that people are more Bayesian in the real world than they appear in some lab experiments. And, the evidence from the lab is mixed; El-Gamal and Grether (1995) find that Bayes' rule is the predominant explanation of their subjects' behavior and Holt and Smith (2007) report similar findings.

[4]For a more thorough discussion of these biases see the references in footnote 1. There is a third, distinct type of non-Bayesian behavior, in which belief revisions take the incorrect sign, rather than simply the wrong magnitude. Overreaction and underreaction, in which revisions have the correct sign but incorrect magnitude, are the focus of this paper.

updating: the Associated Press (AP) college football poll.[5] The AP college football poll is a weekly subjective ranking of the top teams by dozens of journalists who have covered college football for a substantial period of time. Although the data are not economic, they are unique in that they are measures of the evolution of experts' beliefs over time in response to relatively few clearly observed signals, taken from a non-experimental setting. Moreover, due to the richness of the data I observe both over and underreaction to new information by the poll voters, which provides insight into the underlying causes of the different types of errors.

I discuss the data sources in detail in the following Section (2). In Section 3, I discuss my analytical framework and estimation method. I argue that the final, or postseason, individual voter rankings represent the 'true' rankings for each voter-season, and that these true rankings do not change from week to week (for each voter-season), as the final rankings incorporate all relevant information about team qualities and performances for each season. In other words, each voter's pre and mid-season top 25's can be interpreted as predictions of his or her postseason top 25 for that season. This allows use of historical score distributions and empirical final rank frequencies to estimate benchmark Bayesian updated (posterior) rankings. My primary analysis, discussed in Section 4, is based on the first seven weeks of the 2006 season's 64 voters; a total of 11,200 possible prior ranks, each with a corresponding posterior. I first test for differences between the estimated Bayesian posterior rankings' (henceforth I refer to these as the 'estimates'), and the actual, or observed, posterior rankings' (henceforth the 'actuals') distances from truth (the actual final rankings). This tests the accuracy of the estimates. The results indicate that the estimates are more accurate: I reject at the 5% level the null hypothesis that the estimates and actuals are equally accurate for 37.5% of voters in favor of the estimates being more accurate. The null cannot be rejected in favor of the actuals being more accurate for any of the voters. This is taken as evidence of the validity of the estimates. I then construct a variable intended to measure overreaction, defined as excess rank improvement after wins and excess rank decline after losses. The measure of overreaction is statistically affected, both positively and negatively, by several

---

[5]Several other academic studies have used the AP college football poll as a data source, including Lebovic and Sigelman (2001), Goff (1996) and Logan (2007). One major distinction between this work and these studies is that they all rely solely on the aggregate polls, while I use both individual voter and aggregate data. Moreover, no other paper uses the poll to explicitly analyze Bayesian updating.

factors that should have no effect on it under the null hypothesis of Bayesian updating. This implies that voters do in fact systematically underreact in some situations, and overreact in others. The factor that most parsimoniously explains the phenomena is salience: the tendency to overreact increases with the salience, or notability, of the signal and/or prior.

I find that the voters do not appreciate relatively non-salient information, such as home-field advantage. This causes them to over (under) react to home (away) wins, and under (over) react to home losses. The voters also do not respond sufficiently to margin of victory over weak opponents, but do respond rationally to margin of victory of wins over ranked opponents, which are more salient signals. This rather sophisticated behavior reflects the voters' expertise and knowledge of the underlying random processes, enabling them to sometimes be Bayesian. There are other subtleties in the information structure that the voters do not appreciate, however. Perhaps surprisingly, the strongest differences between Bayesian and observed behavior seem to result from non-salient distinctions in the precision of priors. I find that the precision of prior rank is much greater for highly ranked teams.[6] The greater precision implies that responses to losses by top-ranked teams *should* be relatively small: the mean estimated Bayesian rank decline after losses by top 10 teams is 1.3 spots less than the mean estimated rank decline after losses by top 11-25 teams (estimated rank changes of -3.4 and -4.7 for teams ranked 1-10 and 11-25, respectively; this in spite of the fact that top 10 teams have farther to potentially fall). The average actual changes are -6.7 and -4.0. That is, voters reduce top 10 team ranks by 3.3 (= -3.4 - (-6.7)) spots after losses more than they 'should'. Because the voters do not appreciate the non-salient distinctions in strength of prior they overreact to losses by high-ranked teams. Furthermore, the voters in general under-react to signals for low-ranked teams.

If only over or underreaction were observed in the data it might have been attributed to one of the biases or heuristics noted above; since both are observed a more general explanation is required. The positive association between salience of signal and/or prior and overreaction is a more general explanation of behavioral belief updating than many theories discussed in previous literature.

In Section 5 I describe two robustness checks. In the first, I directly compare the actuals to the

---

[6]Nutting (2006) documents this phenomenon in detail, including a reference to a very apt quote made in 1989 by UPI sports editor Fred McMane: "I dont think there are 25 good teams in the country. I think you generally see five good teams, 10 who are fairly good, and after that, who knows?".

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 6 discusses an analysis of the aggregate polls. These data are less appealing because factors that affect individual behavior cannot be accounted for. Still, they are extremely valuable as they allow examination of a wider range of ranks and seasons, since there is greater availability of the aggregate data. The findings from this section support the initial conclusions as well. On average, however, the actual aggregate ranks outperform the Bayesian estimates by the accuracy metric used. Thus, the aggregation of insights and information that voters have that are not taken into account in this analysis outweigh the non-Bayesian mistakes they make. This is partly due to the magnitude of the mistakes not being large, which corroborates the idea that agents make more rational decisions outside the laboratory.

2   The Data

The AP college football poll is a field data source uniquely well suited for analyzing belief updating. 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 obtain other information about each team besides game scores, but this information has a relatively small impact on the rankings, especially on a week-to-week basis.) Moreover, the signal probabilities–the distributions of the scores–are, or should be, common knowledge, since the voters have all observed years of scores. Based on their extensive experience the voters 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. Thus we cannot say, based on observed data, whether or not the individual uses Bayes' rule to update beliefs upon receipt of new information.

While all of the data are technically publicly available, they are not all easily accessible. The current week's aggregate AP poll is ubiquitous in sports publications throughout the college football season. The first poll is taken before the season starts in late August and the final poll occurs after the season ends in early January. The poll is currently voted on by 65 leading college football journalists from throughout the country and different forms of media; the number of voters has varied over the years. 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, but this role ended in 2005. The individual ballots of the AP poll voters are not confidential. The AP 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 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 "James Howell's College Football Scores" (URL as of June 12, 2008: 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 quality 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,[7] and another voter is famous for being the only one to rank the eventual

---

[7]http://sports.espn.go.com/ncf/news/story?id=2663882

championship winner number one early in, and consistently throughout, the 1992 season. In addition, 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 the data only include rankings, rather than a distribution of beliefs regarding a specific variable. In fact, 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 Framework and Estimation

3.1 Defining Truth

The key element to constructing the estimated Bayesian posterior rankings (the *estimates*) is defining truth–the unobserved state of the world that the rankings represent beliefs about. The rankings are never formally defined so this task is not trivial. I define truth, or more accurately, approximate it, as the individual voters' final rankings, for each season. Thus each voter's pre- and midseason rankings can be interpreted as predictions of his/her final rankings for that season. The final rankings approximate the true rankings because they incorporate all potentially available information about "quality of season-long performance" and subjectivity of beliefs about the various teams. The final rankings also incorporate all potentially available information related to "quality". Consequently, if quality is constant throughout the season, each voter's final rankings are the best estimates of the true rankings (for that voter and season), at any point throughout the season.[8,9]

This truth definition does become problematic if voters rank teams on current quality and quality changes throughout the season. Then, for example, if voters' first week rankings were observed to

---

[8]I intentionally avoid specific definitions for the terms season-long performance and quality so as not to impose unnecessary structure. Vaguely, performance refers to the realization of game results, and quality refers to the unobserved team-specific distribution of game results.

[9]The other variable that might appear to plausibly affect rankings is year-to-date (YTD) performance. This variable cannot be what the rankings are based on, however, due to 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. It follows that mid-season polls can also not be based purely on YTD performance. The data bears this out; further information is available from the author upon request.

be very different from their final rankings it would appear that their first week beliefs are very far from "truth", when it would really be just that truth is changing over time. I examine this possibility by testing the hypothesis that average score differences for teams of different final ranks are constant throughout the season.[10] If voters rank teams on current quality, and quality changes, then teams highly ranked in the final poll would have relatively better performances in the later part of the season, on average. This is because teams highly ranked in the final poll would improve on average throughout the year, and teams ranked poorly in the final poll worsen. Please see the Appendix for a theoretical illustration of this phenomenon.

Table 1 presents empirical evidence that the hypothesis cannot be rejected. This indicates that rankings are based primarily on season-long performance, or that team qualities do not change significantly over time, either due to the relatively short season (teams play fewer than 15 games) or perhaps the lack of a hot hand at the team level (Camerer (1989)). Either way the voters' final rankings are valid estimators of their true rankings. Table 1 shows that 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, home teams of final rank 13-25 also perform better in later months versus superior teams of rank 1-12. These data essentially nullify each other. Neither of the other p-values (for games between ranked teams and unranked teams that were ranked in the final poll in one of previous two seasons) are compelling either. There is little reason to lose confidence in the null hypothesis, that score differences are not correlated over time.

I note that truth as defined in this paper is endogenous: it is determined by the voters. Consequently, their belief updating from the second-to-last to last poll is tautologically Bayesian; third-to-last is close to tautologically Bayesian, etc. This endogeneity actually improves the size of the tests; it reduces the likelihood of rejecting the voters being Bayesian when true (a desirable property). Unfortunately, it also reduces the tests' power. To obtain greater statistical power the sample needs to be restricted to the early part of the season. Using early season data also reduces the chance of voters committing the hot-hand fallacy, or falsely inferring trend in team quality changes.[11] On the other

---

[10]I use aggregate ranks for these tests due to lack of historical individual rank data; this should not confound the results.

[11]If voters believed team qualities were changing, they would appear to overreact but would be making mistakes qualitatively distinct from basic misuse of Bayes' rule.

hand, if the sample is restricted to too small a portion of the season, the tests also lose power. I thus limit the analysis simply to the first half (seven weeks) of the season.

One other issue arising from defining truth as the voters' final rankings is that the voters are assumed to 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.[12] This assumption is not accurate if voters are inhibited from expressing their views for fear of embarrassment, or influenced by the beliefs of others for any other non-informational reason. An alternative definition of truth that accounts for these possible issues is the aggregate final rankings. By this definition the true rankings, by team and season, are constant across voters. I conduct the analysis (construction of estimates, and tests of differences between estimates and actuals) using this alternative truth definition, as well as the original definition (individual final rankings), and find that the differences are minimal. The results obtained using the original definition are the preferred estimates, however, and are used whenever the truth definition is not specified.[13]

3.2   Formal Framework

With this discussion in mind, I now specify the voters' objective functions and Bayesian updating process. Let r_i^v denote the true rank of team i for voter v, $i \in \{1, ..., N\}$, $r_i \in \{1, ..., 25, 26+\}$ and $v \in \{1, ..., V\}$, in which N is the number of teams that may be ranked, V is the number of voters and if $r_i = 26+$ the team is unranked. v is 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, the true rank of team indexed 1, may be 10, 20, etc.

Each voter's objective function in week t is to minimize a function of distance between current and true ranks.[14] I make the simple assumption that $E_t(r_i) > E_t(r_j) \rightarrow \widetilde{r}_{i,t} > \widetilde{r}_{j,t}$; in which $\widetilde{r}_{i,t}$ is the

---

[12]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 due to learning; voters know that other voters have superior information about some teams, and they rightly influence one another through their rankings. I discuss this later in the paper.

[13]The data indicate that indeed the voters do not attempt to rank teams as closely to the aggregate ranks as possible. For example, in the first poll of 2006 Ohio State received the majority of first places votes: 35 out of 65. In the second poll, after a strong opening win, it received 39 first place votes. If voters were simply trying to match the aggregate rankings, more than four of them would have switched their first place vote. Moreover, using the aggregate final rankings as the preferred truth definition introduces strategic interaction issues, which could complicate the analysis considerably.

[14]I make the highly defensible assumption that voters do not strategically manipulate their final rankings so as to make their earlier rankings appear to be more accurate.

actual rank of team i in week t. That is, teams are ranked in order of expected rank.

Let s_{ij} be a random variable that is the score difference for the game in which team i plays team j, or 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 with true ranks r_i and r_j results in score s_{ij}.

Let $f_{i,t}(r_i)$ be the (subjective) probability that team i has true rank r_i in week t. (r is indexed by i for clarity in the Bayesian updating formula below.)

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 must change. That is, $\forall k \neq i, j$, the voters update $f_{k,t+1}(r|s_{ij})$. However, 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 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: $f_{10,t+1}(1|s_{10,11})$. Using Bayes' rule, this is equal to 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, $g(s_{10,11})$. The first $g()$ term depends on beliefs about the true rank of team 11, and the second depends on beliefs about the true ranks of both 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:

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

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 then translate these estimated posterior beliefs into rankings we obtain the estimated posteriors.

3.3   Estimation Method

Given this framework, I need estimators of both of the components of 1, the f's and g's, in order to construct the estimates. Please see the Appendix for a discussion of how these distributions are obtained. To summarize, I use smoothed empirical frequencies to estimate both sets of distributions, but am forced to make several assumptions due to data limitations. First, I categorize the support for both the f's and g's; second, I condition on aggregate final rank for the g's; third, I assume all voters have the same f for each prior rank; last, I use the 2006 data to estimate the f's (the same data that the analysis is conducted on). The first and third are approximations and should not introduce any systematic bias. The second may cause the estimated g's to be too 'tight'; if so, this would cause the signals to appear too informative, and the estimates to move too far from the priors. This would bias the results towards findings of underreaction. Of course, using the aggregate final ranks yields the correct g's if aggregate final rank is indeed the correct truth definition. This issue will be kept in mind and discussed as I proceed. The fourth assumptions might cause the estimates to appear overly accurate and affect findings of overreaction (in an unknown way, *ex ante*) if the priors vary year to year and the voters have uncertainty about how this occurs. However, the 2006 aggregate priors appear similar to the 1990-2005 priors;[15] furthermore, estimating priors with current-year data is actually preferable to using historical data if the priors vary year-to-year and the voters are informed about this variation. This is because in this case current-year estimated priors would incorporate information about the relative strength of prior beliefs that would not be reflected in historical-data estimated priors. Taking account of both of these considerations, I do not believe using the current-year data for this purpose is problematic.

The other issue that needs to be addressed before constructing the estimates is the fact that the voters only rank 25 out of approximately 120 teams eligible to be ranked (the number of Division I-A teams). Since most games are between ranked and unranked teams, some objective method of distinguishing among unranked teams is needed. If I treated wins over the best and worst unranked teams equally the estimates would be badly flawed. Fortunately, there are a number of publicly

---

[15]Statistics not reported for brevity but available from the author on request.

observable variables that can be used for this purpose. 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 cardinality of the set of elements $r_{i,t}$ is in to 32, in which $r_{i,t} = 26$ means team i receives at least one vote from others in week t, $r_{i,t} = 27$ 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 unranked 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 actual 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. In other words, it allows the estimates to potentially exactly match the actuals.[16]

---

[16]For 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 relatively little about the new teams in the poll (since they were unranked before) 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. Hence, the estimated rankings can potentially be exactly the same as the actuals.

4   Individual Voter Analysis

4.1   Validity of the Estimates

Before proceeding to the tests of Bayesian updating, it is worthwhile to assess the validity of the estimates. I do this by examining the distances between the estimates and true rankings. I compare these to the analogous distances for the actuals, actual priors and flat priors. To be clear, the actuals are the actual, or observed, posteriors; the actual priors are the voters' rankings prior to the game results, and flat priors are equal rankings for all teams (the average rank). For sufficiently large samples the Bayesian posteriors will on average be closer, by any reasonable measure, to truth than posterior rankings obtained using any other method. Thus, if the estimates and actuals are both approximately Bayesian they will be equally close in distance to the true rankings. If one is less Bayesian its distance from truth will be greater. If either the estimates or actuals is sufficiently flawed then 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. And, if the actual priors are meaningless they will be no closer to truth than the flat priors.

I measure distance from truth by average absolute deviation, by voter: $\frac{1}{n}\sum_{i,t}|\hat{r}^v_{i,t+1} - r^v_i|$; n is the number of observations per voter. I adjust the true rankings as discussed above to account for number of teams, per week and voter, not being in the final poll. Table 3 presents summary statistics for these deviations averaged over all voters. The estimates appear superior to the actuals, and the signal and actual priors appear informative, as they are both lower than the flat prior average deviation. Because observations are correlated across voters by game, I formally test these differences separately by voter. I conduct paired t-tests by individual voter of the difference between the mean absolute deviation from truth for estimated and actual posteriors. The null hypothesis is that the means are the same, and the alternative is that the average deviation for the estimates is smaller. p-values for 58 of the 64 voters are less than 50% (mean estimate deviation less than mean actual deviation). The average p-value 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 actuals incorporate information and personal biases not accounted for by the

estimates, this is strong evidence of the validity of the benchmarks.

4.2   Testing Bayesian Updating

Although the statistics reported above imply that many individual voters are not fully Bayesian, the numbers do not indicate whether or not they have systematic biases, and if so, when the biases come about. In fact, even the voters whose actual posteriors were not significantly less accurate than the estimates very well may not be Bayesian all, or even most, of the time. The two main types of non-Bayesian errors I am interested in testing for are systematic over and underreaction. As mentioned in the introduction there are many well known biases that fall into these two categories. I define overreaction formally (in the context of football rankings) 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 and loss (for simplicity I ignore byes)–and broken out by prior rank categories.

On average the estimated responses to wins are slightly larger than the actuals; since both are positive this is evidence of slight underreaction by the voters. The estimated and actual responses to losses are also similar on average, now indicating slight overreaction, since the magnitude of the actuals is larger. These similar averages, however, mask stark heterogeneity over the various rank groups. In particular, the estimated and actual responses to wins are very different for teams ranked 11-15 and 21-25. The actuals are higher than the estimates for the former group, and lower for the latter, implying that the voters actually overreact to wins by teams ranked 11-15. The responses to losses are substantially different for all rank groups except 21-25. There is apparent strong overreaction to losses for top 10 teams, and moderate underreaction for teams ranked 11-20.

These summary statistics confirm that voters appear to both over and underreact to new information, and to different degrees depending on the context. This apparently contradictory pattern has been recognized in the literature before, but the formal analysis of it has been limited. Barberis and Thaler (2002), without citing any studies, attempt to provide an explanation 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 formally analyze the voters' belief updating processes further, I first define a single variable representing overreaction, $OVER$, as

Definition 4.1. $OVER_{i,t} = \begin{cases} \Delta r_{i,t}^A - \Delta r_{i,t}^E & \text{if } i \text{ wins in } t, \\ \\ \Delta r_{i,t}^E - \Delta r_{i,t}^A & \text{otherwise,} \end{cases}$

in which $\Delta r_{i,t}^j = r_{i,t-1} - r_{i,t}^j$ ($j \in \{A, E\}$, A=actual, E=estimate) is the rank change (improvement) for team i in week t. Of course this means $OVER_{i,t}$ is simply $r_{i,t}^E - r_{i,t}^A$ after wins, and the negative of this expression after losses, but it is useful to think of overreaction in terms of the difference between estimate and actual rank changes. It represents excess rank improvement to 'good' signals (wins; usually signals indicating a team's rank should improve), and excess rank decline in response to 'bad' signals (losses).[17] To control for factors that might affect belief updating under rationality and analyze the determinants of overreaction, I then estimate the following regressions separately for games in which the ranked team wins and loses:

$$OVER_{ijt} = X_{ijt}\beta + \delta_j + WEEK_t * \delta_j + \epsilon_{ijt}. \tag{2}$$

i, j and t denote rank, voter and week, respectively (i-j-t identifies a team-week); δ_j is a voter fixed effect (FE). X is a vector of controls including the following:

1) $HOME$: dummy for team i-j-t playing a home game;

2) $WEEK$: week of the season;

3) $SMARG$: score margin = team i-j-t's points minus opponent's points;

4) $TOP1 - 5$; $TOP6 - 10$; $TOP11 - 15$; $TOP16 - 20$; $TOP21 - 25$: dummies for i in top 1-5, etc.

(constant omitted);

---

[17]This definition is straightforward and easily interpretable; because of its simplicity, however, it does not allow for 'bad' wins (and 'good' losses), which certainly do occur. I experimented with numerous other definitions of overreaction that do account for these types of signals and found that they generally do not result in substantially different estimates.

5) $OPP25$: dummy for i-j-t's opponent ranked in j's top 25 in t;

6) $SM25 = SMARG*OPP25$;

7) $APDEV$: aggregate AP rank of i-j-t minus rank i (if positive j's rank of this team is better than its current aggregate rank; otherwise, j's rank is worse);

8) $RKSD$: standard deviation of rank of i-j-t over voters in t assuming rank of 35 for unranked teams;

9) $RK2YR$: i-j-t ranked by at least one voter in at least one of previous two seasons;

10, 11) ST, REG: dummies for i-j-t being in j's state or census region.

The voter FEs account for voter-specific tendencies to over or underreact. The interaction of the voter FE and week is used to account for possible heterogeneity of ranking definitions. If voters weigh season-performance and quality differently, their responses to games might vary over time (although the final rankings still consistently represent truth).[18] I also show results for alternative specifications with weekly dummies and without the week-voter interaction term. Summary statistics for variables used in the regression analyses are presented in Table 5.

Estimation results, excluding terms with voter FEs, are presented in Tables 6 and 7 with bootstrap standard errors clustered by game.[19] Given the null hypothesis of Bayesian updating, the expected coefficient for most of the variables is zero. Non-zero coefficient estimates are consistent with Bayesian updating for the variables $APDEV$, $RKSD$, and $RK2YR$ 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 almost always significant at at least the 10% level. This means voters tend to overreact to home wins and away losses; they do not appreciate the

---

[18]Specifically, while the magnitudes of all voters' reactions to signals should decrease as the season progresses (and beliefs become more precise), the degree to which the reactions of voters who emphasize performance decrease may be larger. This is because these voters' belief revisions regarding future performance become less important as the number of future games decreases. However, I do not expect this difference to be substantial, as the sample is restricted to the first half of the season and there is a large number of remaining games even after the last week used in the analysis.

[19]Bootstrap standard errors are used because the dependent variable is constructed. Each game, which is defined by its two opponents and the week (and whose definition does not vary across voters), is given an ID number. I cluster errors by game rather than week or voter because residuals are likely to be correlated across voters within games. This is because there is often game-specific information observed by the voters and excluded from this analysis. Estimated standard errors are lower when clustered by any other variable (or not clustered at all). I also exclude the constant so the dummies for all rank groups can be seen clearly.

16

importance of home-field advantage. $SMARG$ is negative for both estimates, and significant for wins. Thus voters are insensitive to margin of victory, but appreciate margin of loss. Voters react more strongly to margin of victory when the opponent is ranked. The rank of the opponent does not have a significant effect on loss responses. The rank group dummies are insignificantly different from zero for wins; but the $TOP21-25$ variable is significantly different from (less than) the Top 11-20 dummies. The top 10 dummies are positive, and strikingly large and significant, for losses.

$APDEV$ takes the expected signs and is significant at the 1% level for most of the estimates. Voters are influenced by their peers; for instance, rank improvements after wins are smaller if the winning team is ranked worse by others. $RKSD$ is not usually significant, though it is consistently negative for wins, indicating that voters respond less strongly to these signals when there is disagreement about the team's rank among peers. $RK2YR$ is also not significant but consistently positive and of high magnitude, with relatively high t-statistics, for wins. This implies that voters are substantially less responsive to wins by teams that have not been in the poll in recent years. Both ST and REG are usually positive, but rarely significant, for wins, providing only weak evidence of 'home bias' in response to new information.[20]

The common feature among these findings is that saliency, or noticeability, is positively associated with the degree to which information is processed. The relationships are sometimes subtle. For instance, home-field advantage is a non-salient characteristic of the signal–it is not emotionally resonant. But it is important; it shifts the score margin distribution to the right (increasing the probability of a win). If voters are unresponsive to its full import, they will overreact to home wins and away losses, which is exactly what I find they do. Score margin might be expected to be more salient since it is more clearly observed. If this were the case, it would exacerbate overreaction as it increases in magnitude. Yet I find that voters underreact more as the margin of victory increases for wins against unranked opponents. Perhaps this is because score margin is not salient when the opponent is perceived to be weak; voters may consider 20 point wins over bad teams equivalent to 60 point wins. The estimates imply, however, that while voters are not responsive to the scores of wins against weaker teams they

---

[20]While the general form of this bias is well known, there is relatively little evidence of the bias also affecting information processing.

17

should be; the best teams do in fact beat teams of all types by the greatest margins. For example, home teams with true rank 1-6 beat teams unranked in the final poll in the last five years by 49+ points 13.5% of the time, while the best unranked teams (those receiving votes in the final poll in last two years) only win this way 4.9% of the time. On the other hand, score margin does not significantly affect overreaction to wins over ranked opponents (F-tests cannot reject $SMARG + SM25 = 0$), which are more salient signals. This is evidence that voters, due to their expertise, are capably of revising beliefs in a sophisticated way when the signal is clearly noticeable.

Even more importantly, many of the findings can be attributed to saliency of the priors, or really a lack thereof. Losses by top 10 teams are the most salient (and alarming, see Viscusi (1997)) observable signal, and the very large coefficient estimates for the $TOP1 - 5$ and $TOP6 - 10$ dummies for losses are consistent with this idea. But losses for all top 25 teams are highly noticeable, so one might struggle to argue that reactions to losses by top 10 teams should be that much greater than reactions to losses by other ranked teams simply due to saliency of the signal. There are significant *relative* differences in the priors, however, which may explain these results. The data show that the priors are much stronger for top 10 teams than other ranked teams, relative to those teams below them. For example, the (raw[21]) mean expected rank of teams ranked in a voter's preseason top 10 is 11.5, and of teams in a voter's week 7 top 10 is 10.9. The corresponding numbers for teams ranked 11-25, and not ranked at all but ranked by at least one other voter, are 25.6 and 22.0, and 28.9 and 27.9. This means that while top 10 teams were on average at least 10 rank spots better than top 11-25 teams, top 11-25 teams were barely better than some unranked teams. Moreover, the mean difference between expected ranks of neighboring top 10 teams in week 1 is 1.3; the mean difference for top 11-25 teams is 0.6. In week 7 these numbers are 1.6 and -0.2.[22] Together, these numbers indicate that priors for high-ranked teams are relatively precise–that the worse a team is ranked, the harder it is to distinguish from similarly ranked teams.

Other things equal, this variation in precision of priors implies that Bayesian responses to losses

---

[21]Expected rank calculated as discussed above, using unsmoothed priors.

[22]The sign of the last number (-0.2) is surprising, it implies that of rank neighbors in the top 11-25, the worse-ranked team had a better expected rank on average in this week. This is likely due to statistical noise and simply be evidence that there is very little distinction between those teams.

by top 10 teams should be relatively small, since prior beliefs about these teams are relatively strong. If voters did not make this non-salient distinction among relative beliefs, however, they would react relatively strongly to losses by high-ranked teams, since these signals are especially unexpected. This is exactly what the summary statistics demonstrate the voters do, and the econometric results confirm. Similarly, voters underreact to wins by low-ranked (21-25) teams. Because the priors for teams ranked 11-25 are very similar, the estimated Bayesian responses to wins by teams ranked 21-25 are large (as they can easily improve greatly) and responses to teams ranked 11-20 small (as they need to compete with top 10 teams to improve substantially). Thus this is additional evidence that the voters do not appreciate the subtle distinctions in prior strength, and that this is a key factor driving non-Bayesian behavior.

I now turn to a brief analysis of voter-level heterogeneity. To do this I use the voter FE estimates from the specifications of the $OVER$ models with weekly dummies. One question of interest is simply is there significant heterogeneity; the answer, unsurprisingly, is yes. F-tests resoundingly reject the hypothesis that the FEs are equal for all of the voters for both the wins and losses models.[23] Other questions include what is the nature of the heterogeneity, and what is the relationship between over-reaction to wins and losses by voter. Figure 1 sheds some light on these questions. It indicates a remarkably sharp positive correlation between overreaction to wins and losses, providing strong evidence that voter tendencies to, at least apparently, over and underreact are general, or similar, across signal types.

Next, I relate the estimates of overreaction to two other observable characteristics of the voters: mean distance between prior and final rankings, and years of poll experience. The distance variable represents accuracy; voters whose prior rankings are very close to their true rankings may have superior information and thus more precise prior distributions, and as a result respond relatively weakly to signals of all kinds. In other words, variation in prior strength may explain variation in voter-level overreaction. If so, this would be seen in a positive correlation between the overreaction FEs (for both

---

[23]These models are estimated with a constant and without a FE for one voter (and without the TOP21-25 variable) to avoid perfect collinearity. This specification is used, rather than the full set of FEs with no constant, because the constant is actually estimated much less precisely than the voter-level effects, and thus when the constant is included the FE standard errors become much smaller.

wins and losses) and the distance variable. Voters with more experience might be expected to be less susceptible to the mistakes made on average discussed above (underreaction to wins and overreaction to losses). This would be seen in a positive correlation between overreaction and experience for wins, and negative for losses. I separately regress the FEs from the wins and losses models on the distance and experience variables to test for these effects but find that all of the estimates are highly insignificant (results omitted for brevity). This implies that voter-level tendencies to overreact, while general across signal types, are not explained by differences in prior strength or experience. A deeper analysis of heterogeneity of tendencies to over and underreact is left for future research.

5   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).[24] The true ranks are unbiased estimators of the Bayesian ranks, but measured with substantial error. The resulting estimates of the effects of signal and prior characteristics on over or underreaction are expected to be much less precise, but qualitatively similar, to those discussed above. Using this method also causes a loss of precision due to loss of sample, because the adjustment to account for teams leaving the poll is more severe. 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 reasonably strong evidence supporting the main results: that belief revisions are excessive when new information is salient and/or strength of prior beliefs is non-salient.

The second robustness check is methodologically much more complex, and qualitatively very different from the other analyses. 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 prior distributions. 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 the earlier conclusions.

---

[24]This methodology is analogous to that used by Amir and Ganzach, as they compare analysts' forecasted earnings changes to actual changes.

5.1   Check 1: Replacing Estimated Posteriors with Truth

Implementing this check is very simple. I again use Definition 4.1 for the variable of interest, overreaction, only replacing Δr^E with Δr^F, in which Δr^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, there are usually 5-10 teams.

Still, the results are strong. The mean overreaction for wins and losses is 0.36 and 1.56, respectively. This reinforces the finding that overreaction to losses is relatively large; the exact numbers are different from those found above due to the greater noisiness of the final rankings as compared to Bayesian posterior rankings. I also estimate equation 2 using the new measure of overreaction as the dependent variable. Results are presented in Table 8; they are similar to, but weaker than, those in Tables 6 and 7. Saliency still sometimes significantly affects information processing ($HOME$ and now $SMARG$ for losses), but not as consistently. Underreaction to wins is still greatest for teams ranked 21-25, and overreaction to losses still greatest for top 10 teams, but the magnitudes are smaller. It is interesting that $APDEV$ is now positive and significant in the 'Wins' model. This reflects voters not being sufficiently influenced by their peers (although the results above showed that voters are influenced to some extent); when a voter ranks a team higher than its aggregate rank the team is likely to be substantially worse than the voter currently believes it to be. Thus, even after wins, teams with poor aggregate ranks are likely to have poor true ranks, and voters appear to overreact when they do not worsen these teams' ranks significantly. Overall, though, these results support the findings of Section 4.

5.2   Check 2: Estimating the Bayesian Priors

As a final robustness check for the individual voter analysis I invert my analytical approach: rather than estimate the Bayesian posteriors and compare these to the actual posterior rankings, I assume the actuals are in fact Bayesian, and estimate the priors that are most consistent with, or best *rationalize*, the data. I compare the expected ranks calculated using the estimated priors to those calculated

using the empirical frequencies of final rank conditional on current (prior) rank throughout the season. Next, I compare the predicted Bayesian posteriors, obtained using the newly estimated priors, to the actual posteriors using the same regression analyses as before. I find that overreaction, defined using predicted Bayesian posteriors, is still positively related to salience (losses, home field, etc.).

5.2.1   Method

To estimate the priors that rationalize the observed behavior I continue to use the Bayesian updating model specified above, 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 priors estimated as above, 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 prior for each week, seven priors total, since the priors change week to week under rationality. Formally, I search for

$$f_t^* = \underset{f_t}{arg min} \sum_i \sum_j |\hat{r_{ijt}}(f_t) - \widetilde{r}_{ijt}|, \quad t = 1, ..., 7, \tag{3}$$

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 $\widetilde{r}$ is actual posterior. I use sum of absolute deviations as a distance metric to minimize sensitivity to outliers.

A key step here is the choice of priors to search over. The set of possibilities is infinite. The problem is also made more complicated as the term 'prior' as I have used it actually refers to a joint distribution of order statistics. Specifically, each prior (f) is the joint distribution of categorized final rank conditional on categorized prior rank for all 119 (the number of Division I-A) teams. As described in the Appendix, I categorize final ranks for all teams (1-6, 7-12, 13-18, 19-25 and unranked), and prior ranks for unranked teams (seven categories), so that there are 32 (=25+7) prior rank categories and five final rank categories. This means that each f is a 32 x 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 final rank category (6, 6, 6, 7 and 94(=119-25)).

In order to search over these joint distributions I must first generate a set of matrices with the features described above, and consistent with the idea that teams are ranked in order of expected rank. I do this by assuming teams are ranked on a jointly normal latent variable, with zero covariance across teams. For concreteness I refer to this variable here as quality. I equally space the means of beliefs about the quality variable across the unit interval (0, 0.0085, ..., 0.9915, 1) for all 119 teams. Next, rather than assuming that each quality variable has the same variance and searching for the optimal value of this variance, I assume that variances are constant only within categories of prior ranks. For this purpose I use the final rank categories for top 25 teams (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 unranked 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 prior rank categories with (potentially) distinct variances.

I permute five different variances (0.002, 0.006, 0.010, 0.014, 0.018) over the six rank groups. This leads to the generation of a set of $5^6 = 15,625$ joint prior distributions to search over, each of which I estimate through simulation of the latent jointly normal quality variables (10,000 runs). This method of determining the prior set that I search over is fairly arbitrary. It serves the purpose at hand, 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 statistic distributions to search over. This method yields this large and varied set.

5.2.2 Results

Table 9 illustrates some properties of the estimated Bayesian priors, which I henceforth refer to as simply the Bayesian priors, in comparison with the analogous empirical ones. It shows, first of all, that the Bayesian prior expected ranks are fairly uniformly spaced relative to the actuals, especially from week three forward. The Bayesian prior expected ranks for teams ranked 13-18 and 19-25 are

approximately 17-19 and 23-25, respectively, for weeks 3-6. The actual expected rank for teams in both of these rank groups is around 25 for this time period. Thus the voters act as if beliefs distinguishing teams ranked 13-25 are relatively precise; teams ranked 13-18 are thought to be substantially better than those ranked 19-25. In reality, the true ranks of these two groups of teams are basically indistinguishable.

The other important phenomenon revealed in this table relates to teams with prior rank 1-6. The actual expected rank of these teams declines steadily over time, as one would expect given that information about the teams is accumulated over time and beliefs about their true ranks become more precise. The Bayesian prior expected ranks, however, start quite low, and jump up in weeks three, six and seven. These were the only weeks when teams with aggregate rank 1-6 lost games. This implies that voters acted as if their priors for top-ranked teams were weak when they lost, and strong when they won, which is completely consistent with the finding that voters overreact to losses by top-ranked teams. Note also that the gap between expected rank of top 1-6 and 7-12 teams is much larger for the actuals than estimates, which is also evidence that voters did not appreciate how strong their priors for top 1-6 teams were.

Finally, I again conduct the overreaction regression analysis discussed above. I continue to use Definition 3.1 for the dependent variable, overreaction, now using the predicted posterior ranks as the estimated Bayesian posteriors. Overreaction is 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: mean redefined $OVER$ is 0.19 and 0.97 for wins and losses, respectively. Even though I have fit the priors to the data there is still evidence of overreaction to more salient signals (losses). The values are all positive because the predicted rank changes regress to the mean 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.

6 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 are worth exploring for various reasons. First, they allow priors to be estimated using historical final rank frequencies. Second, I can conduct the analyses for multiple seasons, and confirm that any trends found are not confined to 2006 for some reason. Third, the aggregate polls include rankings for a larger set of teams–30-50 for most weeks, since since aggregate point totals for all teams receiving votes are reported, 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 I have already analyzed it to see how using 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 method in which I do not smooth priors 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. These numbers imply 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 -0.18 and -0.09 for wins and losses, respectively. Overreaction is still greater to losses than wins, but on average it is close to zero. These numbers are misleading, however, as is seen when the statistic is again split out by initial rank group. Mean overreaction to losses for top 10, 11-25 and 26+ teams are 3.06, -2.15 and 0.24, respectively. Overreaction is still strongest on average in response to losses by top-ranked teams. But there is strong underreaction to losses by middle ranked teams. This is 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 for an individual voter could have a maximum rank decline of five spots; in the aggregate analysis, its rank decline can be up to 15 spots. It is worth noting that because these mistakes occur in opposite directions they may nullify each other to some extent. For example, if a top 10 team loses it may fall too far in the rankings and become a top 11-25 team. But if it wins its next game it may rise too far, partly making up for the previous excessive decline.

I also confirm that priors are relatively stronger for top 10 teams. Expected ranks in week 1 for top 10, top 11-25 and top 26-35 teams are 12.5, 27.7, and 32.7. Numbers are very similar in week 4. Thus, even taking into consideration the censored nature of the data, the difference between expected ranks of top-ranked and middle-ranked teams is again much greater than the difference between middle- and low-ranked teams. It is very important to confirm that this trend was not confined merely to the 2006 season.

Regression analysis also yields results similar to those found using the individual voter data. I again estimate equation 2, now dropping voter-specific variables, adding year fixed effects and using broader rank group dummies; $TOP10$, $TOP11-25$ and $TOP26p$ (26+). Results are presented in Table 10. For wins, voters, even in the aggregate, fail to fully appreciate home-field advantage and score margin

in general, but again the voters do respond to score margin when it is salient, i.e. games in which the opponent is ranked in the top 25. Regarding losses, the importance of home is not recognized, while there is no evidence that score is over-reacted to. And while $TOP10$ is not significant, it is significantly different from $TOP11-25$ at the 1% level. Again, we see that voters overreact to losses when the priors are strong and do the opposite when weak. Voters also overreact to score margin after losses when the opponent is in the top 25.

As a robustness check for these results, I also construct estimates using unsmoothed priors and estimate the regressions. I can do this for the aggregate, but not the individual, analysis because of the availability of historical aggregate rank data. Still, as noted above, the raw historical final rank frequencies are very noisy. Using this method, 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. Mean overreaction to wins is 0.03 and 0.43 to losses. Mean overreaction to losses by top 10, 11-25 and 26+ teams is 2.81, -2.33, and 1.4, respectively. The regression estimates are similar to those found using smoothed priors, indicating that the results in general are not sensitive to the smoothing method.

7   Concluding Remarks

This study has compiled extensive evidence of both Bayesian updating and over and underreaction to new information by experts in a real-world context. The errors are usually not large in an absolute sense, and sometimes nullify each other, but they are systematic and statistically significant. The errors cannot be explained by any single known bias, such as anchoring or recency, since they point in opposite directions. A more general factor that appears to explain the data is salience: voters do not pay full heed to non-salient new information and subtle distinctions in strength of priors. Namely, voters improve teams' rankings excessively after wins at home. The voters worsen rankings excessively after losses on the road, by large margins, and in general, and especially, for top-ranked teams. Decisive wins against unranked teams are unappreciated by voters. Individuals who over (under) react in general to wins do the same to losses, but this is not because they have different prior precisions.

Hopefully, these results will strengthen economists' confidence in experimental findings of non-Bayesian behavior, and enhance understanding of when the various types of non-Bayesian mistakes are likely to come about. Using experiments to confirm the relationship between salience and overreaction, analyzing individual-level heterogeneity at a deeper level, and applying these results to the study of real-world economic phenomena are important directions for future research.

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," *NBER Working Paper*.

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.

DELAVANDE, A. (2008): "Measuring revisions to subjective expectations," *Journal of Risk and Uncertainty*, 36(1), 43–82.

DELLAVIGNA, S. (2007): "Psychology and Economics: Evidence from the Field," Working Paper.

EL-GAMAL, M., AND D. GRETHER (1995): "Are People Bayesian? Uncovering Behavioral Strategies.," *Journal of the American Statistical Association*, 90(432).

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.

HOLT, C., AND A. M. SMITH (2007): "An Update on Bayesian Updating," Working Paper.

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," *NBER Working Paper.*

NUTTING, A. (2006): "The Marginal Accuracy of Weekly College Football Rankings: Are There Diminishing Returns?," Working Paper.

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.

VISCUSI, W. (1985): "Are Individuals Bayesian Decision Makers?," *American Economic Review*, 75(2), 381–385.

VISCUSI, W. (1997): "Alarmist Decisions with Divergent Risk Information," *The Economic Journal*, 445, 1657–1670.

A   Estimation

A.1   Testing Constant Team Qualities

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 change over time so that $X_{i,2} = X_{i,1} + \epsilon_i$, in which ϵ_i is i.i.d. with mean 0 and positive variance.

Let $\widetilde{r}_i$ denote team i's final rank. Voters rank teams on current quality so $\widetilde{r}_1 = 1 \leftrightarrow s^2 > 0$ (team 1, or the home team, finishes with final rank 1 if and only if it wins the second game; 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

The natural way to estimate the score distributions (the g's) 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. The other issue complicating the estimation of the g's 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 AP Top 25 in its current form began.[25] 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 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 divide the true top 25 into four categories; 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 category. Then, I divide the score distribution into categories of size 7 (with upper and lower bounds of plus/minus 49+). I construct estimates using both the raw frequencies of scores in each bucket for games between teams in each rank group, and smoothed frequencies obtained using what is essentially a uniform kernel with bandwidth two, and the differences are minimal. The smoothed estimates are referred to by default in the body of the paper, and the smoothing method is explained as follows.

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 $c^i_{j,k}$ denote the historical count of games with score margins in category $i \in \{1, ..., 17\}$ for games between home team of rank group j and away team of rank k. j and k are henceforth suppressed. For $i \in \{3, ..., 6, 12, ..., 15\}$ let $\widetilde{c}^i = \frac{1}{5}\Sigma^{i+2}_{\hat{i}=i-2}c^{\hat{i}}$. For $i \in \{1, 2\}$ let $\widetilde{c}^i = \frac{1}{3+I(i=2)}\Sigma^{i+2}_{\hat{i}=1}c^{\hat{i}}$, in which $I(i=2) = 1$ if $i = 2$, else $I(i=2) = 0$. For $i \in \{16, 17\}$ let $\widetilde{c}^i = \frac{1}{3+I(i=16)}\Sigma^{18}_{\hat{i}=i-2}c^{\hat{i}}$. $\widetilde{c}^i = c^i$ if $i = 0$. Let $g(s^i_{jk})$ 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 $\hat{g}(s^i_{jk}) = \frac{\widetilde{c}^i}{\Sigma^{17}_{\hat{i}=1}\widetilde{c}^{\hat{i}}}$.

---

[25]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.

A.3 Prior Distributions

I estimate the prior distributions using the empirical frequencies of true (final) rank conditioned on current rank.[26] I assume that voters have the same belief distributions conditional on prior rank (e.g., each voter's number one team in week one has the same probability of being the true number one, two, etc.). Thus 25 distributions need to be estimated (one for each ranked team), for each week. Recall that by assumption the voters have rational expectations: $\tilde{r}_{i,t} < \tilde{r}_{j,t} \rightarrow E_t(r_i) \leq E_t(r_j), \forall i, j, t$. Table 2 shows that due to the limited sample size, monotonicity of expected rank is violated in the raw frequencies.[27]

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 in rank group 1-6 as the frequency of teams with prior ranks 1-3 finishing 1-6. The practical question then is how many neighboring ranks should be used for this purpose. An extreme approach would be to use the minimum number to obtain monotonicity of expected rank over all 25 prior ranks), but this can lead to over-smoothing. 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.

Formally, 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. The priors for the unranked teams are estimated as the raw empirical frequencies of final rank. Let $n_{jt,k}^v$ 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

---

[26]It is reasonable to believe that other observable factors, such as teams' ranks in the previous week, affect the priors. I ignore these factors for simplicity.

[27]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.

team is currently unranked), $f_{i,t}(k|r_{it} \in j)$ is estimated as $\frac{\Sigma_v n_{jt,k}^v}{\Sigma_k \Sigma_v n_{jt,k}^v}$.

If $j \leq 25$ 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 used, I estimate the probability team i with prior rank j has true rank k (in period t), $f_{i,t}(k|r_{it} = j)$, as $\frac{\sum\limits_{\hat{j}=j-b:j+b} \Sigma_v n_{\hat{j}t,k}^v}{\sum\limits_{\hat{j}=j-b:j+b} \Sigma_k \Sigma_v n_{\hat{j}t,k}^v}$. 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 (each voter's expected number of teams finishing in rank group 1, ranks 1-6, equals six, etc.).

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 of expected rank is violated after smoothing, I simply average the estimated priors for j and $j+1$ (and $j + 2$, etc., as necessary):

$$\hat{f}_{i,t}(k|r_{it} = j) = \hat{f}_{i,t}(k|r_{it} = j + 1) = \frac{\sum\limits_{j,j+1} \left[\sum\limits_{\hat{j}=j-b:j+b} \Sigma_v n_{\hat{j}t,k}^v \right]}{\sum\limits_{j,j+1} \left[\sum\limits_{\hat{j}=j-b:j+b} \Sigma_k \Sigma_v n_{\hat{j}t,k}^v \right]}. \tag{4}$$

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.

B Tables and Figures

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

Prior	Pr(r ∈ [1,6])	Pr(r ∈ [7,12])	Pr(r ∈ 13,18])	Pr(r ∈ [19,25])	Pr(r ∈ [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 (3.5 for [1,6], 9.5 for [7,12], etc.) and 35 for unranked category.

Table 3: Absolute Deviations from Postseason Ranks

	Estimates	Actuals	Priors	FlatPrior
mean	3.39	3.61	3.73	5.05
sd	3.19	3.32	3.43	2.70

N = 10,893

Table 4: Rank Change (Prior Rank - Posterior Rank) Means and Standard Deviations

Prior Rank	Wins		Losses		Byes	
	Actuals	Estimates	Actuals	Estimates	Actuals	Estimates
1-5	0.07	-0.12	-6.63	-2.83	0.10	-0.91
	(1.23)	(1.27)	(2.79)	(2.35)	(0.82)	(1.33)
6-10	0.35	0.45	-6.80	-3.94	-0.06	-1.92
	(1.91)	(1.92)	(4.00)	(3.29)	(1.14)	(2.66)
11-15	1.34	0.32	-6.59	-8.01	0.07	-5.50
	(2.41)	(2.49)	(3.20)	(2.65)	(1.45)	(1.68)
16-20	1.85	1.44	-4.52	-5.33	0.81	-2.59
	(2.52)	(2.95)	(2.24)	(1.76)	(1.83)	(1.94)
21-25	2.35	4.63	-0.65	-0.48	1.14	2.18
	(2.37)	(3.91)	(1.19)	(1.39)	(1.40)	(1.68)
Total	1.15	1.31	-4.76	-4.32	0.43	-1.91
	(2.29)	(3.15)	(3.63)	(3.51)	(1.49)	(3.18)

N = 8028, 2008, 857 (Wins, Losses, Byes); Prior Rank = actual, observed individual voter rank, weeks 1-7 of 2006 season; Posterior Rank = following week rank.

Table 5: Descriptive Statistics of Variables Used for Econometric Analysis

|	Variable	Mean	Std. Dev.	Min	Max
Wins	OVER	-0.167	2.958	-15	15
n = 8028	HOME	0.670	0.470	0	1
	SMARG	23.634	14.697	1	62
	OPP25	0.159	0.366	0	1
	SM25	2.347	6.429	0	42
	APDEV	0.386	3.117	-13	18
	RKSD	3.241	1.777	0	8
	RK2YR	0.975	0.156	0	1
	ST	0.032	0.177	0	1
	REG	0.099	0.298	0	1
Losses	OVER	0.443	3.296	-13	16
n = 2008	HOME	0.407	0.491	0	1
	SMARG	-12.269	8.674	-42	-1
	OPP25	0.635	0.481	0	1
	SM25	-9.384	9.961	-42	0
	APDEV	0.674	3.553	-12	17
	RKSD	3.452	1.414	1	7
	RK2YR	0.965	0.183	0	1
	ST	0.034	0.182	0	1
	REG	0.099	0.299	0	1

Variables defined in body text.
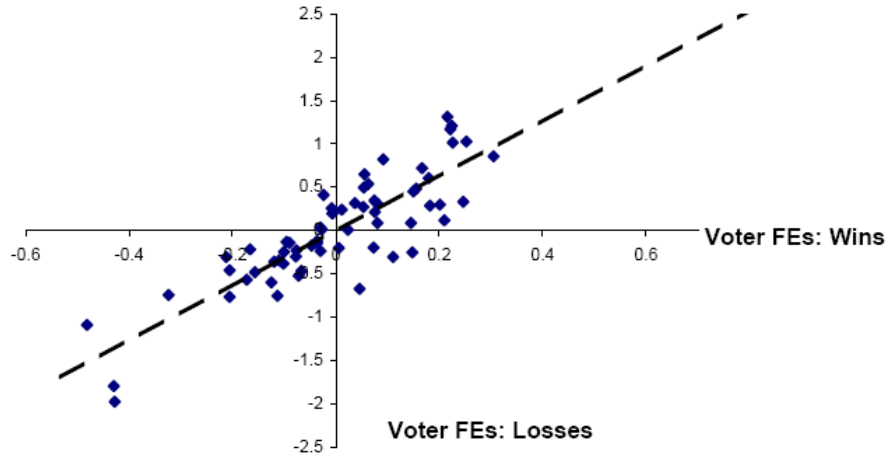
Figure 1: Scatterplot of De-Meaned Voter Fixed Effect Estimates with Trend Line

Table 6: Estimation Results: Wins

	A	B	C	D	E	F
HOME	1.682***	1.669***	1.631***	1.487***	1.495***	1.424***
	(0.333)	(0.356)	(0.352)	(0.370)	(0.339)	(0.357)
SMARG	-0.056***	-0.088***	-0.055***	-0.057***	-0.086***	-0.056***
	(0.010)	(0.011)	(0.010)	(0.012)	(0.009)	(0.011)
TOP1-5	-0.060	0.944	0.068	-0.358	0.496	-0.005
	(1.282)	(1.522)	(1.235)	(1.505)	(1.566)	(1.471)
TOP6-10	-0.246	0.800	-0.071	-0.596	0.343	-0.196
	(1.278)	(1.479)	(1.247)	(1.562)	(1.522)	(1.570)
TOP11-15	1.221	2.369	1.370	0.960	1.893	1.336
	(1.292)	(1.582)	(1.311)	(1.609)	(1.606)	(1.587)
TOP16-20	0.901	2.082	1.039	0.694	1.675	1.059
	(1.384)	(1.714)	(1.289)	(1.620)	(1.593)	(1.623)
TOP21-25	-1.555	-0.667	-1.394	-1.952	-1.150	-1.567
	(1.382)	(1.718)	(1.296)	(1.627)	(1.491)	(1.621)
OPP25	-1.541**	-1.771***	-1.906***	-1.410*	-1.688**	-1.851***
	(0.761)	(0.569)	(0.709)	(0.786)	(0.752)	(0.614)
SM25	0.063	0.078***	0.074*	0.058	0.070*	0.071*
	(0.044)	(0.029)	(0.042)	(0.056)	(0.037)	(0.041)
APDEV	-0.122***	-0.110***	-0.124***	-0.127***	-0.120***	-0.129***
	(0.030)	(0.027)	(0.028)	(0.034)	(0.034)	(0.020)
RKSD	-0.240	-0.259	-0.262**	-0.180	-0.177	-0.204
	(0.158)	(0.191)	(0.122)	(0.159)	(0.164)	(0.149)
RK2YR	1.483	1.387	1.409	1.467	1.437	1.386
	(1.182)	(1.540)	(1.143)	(1.332)	(1.562)	(1.272)
ST	0.228	0.143	0.224**	0.110	0.016	0.107
	(0.140)	(0.118)	(0.091)	(0.125)	(0.114)	(0.137)
REG	0.087	0.024	0.092	0.071	-0.017	0.077
	(0.091)	(0.104)	(0.102)	(0.095)	(0.090)	(0.090)
Truth: Ind.	✓	✓	✓			
Smoothed $g()$	✓		✓	✓		✓
Week FE			✓			✓
R-sq	0.294	0.336	0.304	0.280	0.315	0.292
N	8028	8028	8028	8028	8028	8028

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

Bootstrap standard errors clustered by game in parentheses. "Truth: Ind." denotes individual final rankings used as true rankings; otherwise, aggregate final rankings. "Smoothed $g()$" denotes smoothed score distributions used; otherwise, raw score distributions. "Week FE" denotes weekly dummies; otherwise, categorical week variable interacted with voter FE.

Table 7: Estimation Results: Losses

	A	B	C	D	E	F
HOME	-1.430*	-1.373*	-1.523*	-1.109*	-1.255**	-1.074
	(0.781)	(0.797)	(0.882)	(0.584)	(0.523)	(0.673)
SMARG	-0.057	-0.040	-0.023	-0.038	-0.030	-0.012
	(0.059)	(0.049)	(0.038)	(0.042)	(0.054)	(0.037)
TOP1-5	5.993**	6.682**	7.658***	4.532**	6.108**	6.002***
	(2.476)	(2.637)	(2.146)	(2.121)	(2.427)	(1.623)
TOP6-10	4.755**	4.554**	6.182***	3.367**	4.248**	4.660***
	(2.044)	(2.045)	(1.969)	(1.487)	(1.891)	(1.552)
TOP11-15	0.296	1.317	1.647	-0.557	0.946	0.794
	(1.969)	(2.240)	(2.024)	(1.586)	(1.846)	(1.826)
TOP16-20	0.107	1.303	1.698	-0.957	0.658	0.587
	(1.765)	(2.299)	(1.958)	(1.560)	(1.814)	(1.795)
TOP21-25	0.884	1.938	2.560	-0.176	1.150	1.480
	(1.837)	(2.402)	(1.989)	(1.621)	(1.887)	(1.870)
OPP25	-0.620	0.381	-1.311	-0.151	0.343	-0.913
	(1.017)	(0.861)	(0.888)	(0.765)	(0.781)	(0.759)
SM25	0.040	0.098	-0.019	0.029	0.073	-0.023
	(0.070)	(0.069)	(0.051)	(0.059)	(0.070)	(0.050)
APDEV	0.097**	0.107***	0.108***	0.131***	0.129***	0.151***
	(0.044)	(0.039)	(0.036)	(0.037)	(0.045)	(0.033)
RKSD	0.200	0.088	0.157	0.071	0.094	0.011
	(0.269)	(0.277)	(0.330)	(0.204)	(0.249)	(0.268)
RK2YR	-1.141	-0.952	-1.206	-0.433	-0.844	-0.398
	(0.951)	(1.024)	(1.064)	(0.733)	(1.014)	(0.843)
ST	0.048	-0.082	0.080	0.112	-0.030	0.110
	(0.286)	(0.302)	(0.304)	(0.332)	(0.368)	(0.306)
REG	0.107	-0.029	0.051	0.182	0.028	0.105
	(0.270)	(0.219)	(0.220)	(0.248)	(0.270)	(0.198)
Truth: Ind.	✓	✓	✓			
Smoothed $g()$	✓		✓	✓		✓
Week FE			✓			✓
R-sq	0.472	0.400	0.485	0.473	0.426	0.481
N	2008	2008	2008	2008	2008	2008

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

Bootstrap standard errors clustered by game in parentheses. "Truth: Ind." denotes individual final rankings used as true rankings; otherwise, aggregate final rankings. "Smoothed $g()$" denotes smoothed score distributions used; otherwise, raw score distributions. "Week FE" denotes weekly dummies; otherwise, categorical week variable interacted with voter FE.

Table 8: Estimation Results: Robustness Check 1

	Wins	Losses
HOME	-0.835	-3.016***
	(0.711)	(1.137)
SMARG	-0.036	-0.158*
	(0.030)	(0.084)
TOP1-5	0.501	2.910
	(2.219)	(3.253)
TOP6-10	-0.996	3.442
	(2.442)	(3.078)
TOP11-15	-0.660	2.174
	(2.360)	(3.005)
TOP16-20	-2.357	2.425
	(2.416)	(2.944)
TOP21-25	-4.558*	2.099
	(2.416)	(3.044)
OPP25	1.246	3.456**
	(1.975)	(1.523)
SM25	-0.042	0.274***
	(0.114)	(0.103)
APDEV	0.133**	-0.054
	(0.065)	(0.068)
RKSD	-0.076	-0.302
	(0.347)	(0.429)
RK2YR	3.081	-1.632
	(1.896)	(1.819)
ST	-0.054	-0.340
	(0.203)	(0.285)
REG	-0.191	0.067
	(0.181)	(0.212)
R-sq	0.214	0.433
N	8028	2008

Significance levels :   * : 10%   ** : 5%   *** : 1%.
Bootstrap standard errors clustered by game in parentheses.

Table 9: Estimate/Actual Expected Final Rank Comparison

	Prior Rk	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7
Estimates	1-6	6.62	6.72	9.70	6.31	6.92	9.79	8.79
	7-12	12.86	11.78	13.90	10.93	13.20	14.71	12.45
	13-18	16.74	18.53	18.24	16.90	16.82	18.45	19.24
	19-25	23.37	24.33	23.27	23.02	22.86	24.46	25.44
Actuals (Raw)	1-6	10.21	10.08	8.57	6.85	6.77	6.56	7.29
	7-12	15.90	17.06	16.72	16.41	16.57	16.73	18.21
	13-18	23.37	24.25	25.01	25.57	24.66	25.51	23.33
	19-25	28.83	27.41	26.35	25.49	24.84	24.21	20.85

Notes: Estimates calculated using 'Bayesian priors' (see body text for definition); Actuals calculated using 2006 empirical frequencies of final ranks; expected ranks based on mid-point ranks for each rank category (3.5, 9.5, 15.5, 22) and 35 for unranked.

Table 10: Estimation results: Aggregate Rankings

	Wins		Losses	
	A	B	C	D
HOME	1.472***	1.504**	-1.499***	-1.757**
	(0.422)	(0.669)	(0.543)	(0.791)
SMARG	-0.067***	-0.076***	0.093	0.158*
	(0.012)	(0.018)	(0.065)	(0.085)
OPP25	-0.255	0.538	-0.575	-1.198
	(0.764)	(1.191)	(1.139)	(1.441)
SM25	0.103	0.070	-0.133**	-0.166*
	(0.077)	(0.061)	(0.063)	(0.088)
TOP10	0.017	0.785	2.090	2.239
	(0.621)	(0.968)	(1.622)	(1.798)
TOP11-25	0.253	1.493	-3.716**	-3.445*
	(0.674)	(1.042)	(1.534)	(1.886)
TOP26p	-0.884	-1.517*	-0.668	1.159
	(0.642)	(0.843)	(1.242)	(1.449)
Smoothed $f()$	✓		✓	
R-sq	0.140	0.122	0.328	0.263
N	352	352	104	104

Significance levels :   * : 10%   ** : 5%   *** : 1%.

Bootstrap standard errors in parentheses. "Smoothed $f()$" denotes use of smoothed empirical final rank frequencies for prior distributions.