Federal Reserve Bank of New York
Staff Reports

Bayesian Social Learning, Conformity, and Stubbornness: Evidence from the AP Top 25

Daniel F. Stone
Basit Zafar

Staff Report no. 453
June 2010

This paper presents preliminary findings and is being distributed to economists and other interested readers solely to stimulate discussion and elicit comments. The views expressed in this paper are those of the authors and are not necessarily reflective of views the Federal Reserve Bank of New York or the Federal Reserve System. Any errors or omissions are the responsibility of the authors.

Bayesian Social Learning, Conformity, and Stubbornness: Evidence from the AP Top 25

Daniel F. Stone and Basit Zafar

Federal Reserve Bank of New York Staff Reports, no. 453
June 2010
JEL classification: D80, D83, D84

Abstract

The recent nonexperimental literature on social learning focuses on showing that observational learning exists, that is, individuals do indeed draw inferences by observing the actions of others. We take this literature a step further by analyzing whether individuals are Bayesian social learners. We use data from the Associated Press (AP) U.S. College Football Poll, a weekly subjective ranking of the top twenty-five teams. The voters' aggregate rankings are available each week prior to when voters have to update their individual rankings, so voters can potentially learn from their peers. We find that peer rankings: 1) are informative, as conditioning on them improves the accuracy of our estimated Bayesian posterior rankings in a nontrivial way, and 2) influence the way voters adjust their rankings, but the influence is less than the Bayesian amount. Voters' revisions are closer to Bayesian when the ranked team loses as compared to when it wins, which we attribute to losses being less ambiguous and more salient signals. We find evidence of significant voter heterogeneity, and that voters are less responsive to peer rankings after they have been on the poll a few years. We interpret the data to imply that reputation motives cause voters to "conform," but not enough to overcome the overall tendency to underreact to social information, that is, to be "stubborn."

Key words: belief, Bayesian updating, social learning, conformity, herding, peers

Stone: Oregon State University (e-mail: dan.stone@oregonstate.edu). Zafar: Federal Reserve Bank of New York (e-mail: basit.zafar@ny.frb.org). The authors thank Da Lin for excellent research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

1 Introduction

It has become widely recognized that observational learning–individuals drawing inferences on the private information of others by observing their actions–affect a large class of economic phenomena.[1] To give just a few examples, there is evidence voters learn about politicians, farmers learn technologies, and homeowners learn about mortgage defaults from their peers.[2] Two recent papers that discuss especially clear results are Cai, Chen, and Fang (2009) and Moretti (2008). The former analyze a field experiment in which restaurant consumers are randomly given information on top-selling items. The authors find that this information has a considerable effect on purchase decisions.[3] Moretti (2008) analyzes non-experimental data on movie box office sales, finding several results consistent with consumers learning about movie qualities by observing whether box office sales were above or below expectations.[4]

This research provides convincing evidence on the existence and importance of social learning, which is non-trivial, as cleanly distinguishing social learning from other factors that cause similar observed behavior is usually difficult. However, the empirical literature on social learning is agnostic on the normative *degree* of social learning–that is, whether individuals are influenced by others excessively, insufficiently, or just the right amount. We take this literature a step further by analyzing whether individuals are rational, i.e., Bayesian–social learners. To do this we use a rich, real-world data source: the voter ballots of the AP Top 25 U.S. college football poll for the 2006-08 seasons. The poll is a subjective, weekly ranking of the top 25 (out of more than 100) teams, voted on by over 60 experienced sports journalists, giving us over 30,000 observations. The data source is particularly well-suited for the analysis of how individuals' beliefs respond to social information for several reasons. First, it allows us to observe the evolution of beliefs of individuals over time in response to observable signals (game scores). Second,

[1] We use the terms social learning and observational learning interchangeably in this paper. In other contexts, observational learning is a strict subset of social learning, as the latter may also include direct communication. In our context the terms are equivalent.

[2] See Knight and Schiff (2007), Conley and Udry (2010) and Cohen-Cole and Duygan-Bump (2008).

[3] The authors analyze the effects of providing a display on restaurant tables with information on the most popular dishes. They distinguish information effects from saliency effects by comparing outcomes when the display says the dishes are popular as compared to conditions in which the display simply names selected dishes. The authors caution that they cannot distinguish between learning and conformity, which they imply is simply the desire to be similar to others, but argue that conformity is unlikely to drive their results due to the nature of the restaurant context.

[4] The author uses the number of theaters a movie is showing in to proxy expectations of movie "quality", and shows that when sales are relatively high in the first week of release, given the number of theaters, sales decline at a relatively slow rate (and sales decline relatively quickly when sales are initially below expectations). This indicates that some consumers who attend movies in later weeks drew inferences from the initial weeks' sales results. When initial week sales are high or low for reasons that do not provide information on movie quality, such as weather shocks, the patterns do not hold, so the results are unlikely to all be explained by factors other than social learning, such as the possibility that movie-goers simply have correlated private information.

aggregated poll results are widely available each week of each season, so voters can observe their peers' rankings before updating their rankings. Third, the data source allows us to identify deviations from Bayesian social learning–a task that is all but impossible in most empirical settings–by using each voter's final rankings, for each season, to proxy her/his true rankings for that season. This assumption is questionable, but we believe it is quite natural and holds up under scrutiny. Intuitively, the final rankings reflect all information that will ever be available on team qualities and performances for that season, and idiosyncratic voter preferences and biases. So it makes sense to think of the voters as trying to "match" their current rankings to final rankings each week throughout the season. We discuss this assumption in detail in section 3.

Section 3 also discusses our empirical approach, which is an adaptation of that used by Stone (2009). It involves multiple steps but is conceptually straightforward. We first directly estimate each voter's Bayesian posterior rankings, by week and season, using empirical distributions to estimate voters' prior and game score distributions. The estimated posteriors are conditioned both on game scores and the aggregate rankings (i.e., the rankings of other voters), which we refer to as the social information. If voters had completely idiosyncratic tastes regarding true rankings, the aggregate rankings would be uninformative (with respect to the voters updating their individual ballots) and the social information would not affect our estimated posteriors; if voters had similar tastes and heterogeneous information, the aggregate rankings would be informative. In other words, to be clear, our empirical method allows for the possibility of the aggregate rankings being informative, but does not assume they are–we let the data speak for itself on this issue, indirectly via our estimated posteriors.[5]

Our next step is to assess the validity of our estimates. We find that our estimated posteriors match the voters' own final rankings better than their own posteriors do. This is evidence that our estimates are "more Bayesian" than the observed posteriors, providing support for the validity of our estimates. This allows us to use our estimates to test for systematic non-Bayesian behavior, despite the fact that our estimates are clearly based on a limited subset of the relevant information actually available to voters. We also find strong evidence that taking

[5] We do expect *a priori* that the voters have similar tastes in rankings. To illustrate with an extreme example, it is natural to think all voters would rank an undefeated team better than a winless team. Consequently, assuming that voters have heterogeneous information on the characteristics of the various teams, it is natural to think voters can learn from other each other about how best to rank the teams.

account of social information does in fact make our estimated posterior rankings more accurate, implying voters' tastes are correlated and information is heterogeneous, so voters indeed can learn from their peers.

We then use straightforward regressions to test the null hypothesis that the voters are Bayesian social learners. We find considerable evidence supporting rejection of the null. In particular, voters underreact to winning teams having a better aggregate rank: this should cause the ranks for these teams to improve by around 3 spots, but voters only improve them by around 1.5 spots (as compared to winning teams with similar aggregate ranks). Also, voters underreact to winning top 15 teams having a worse aggregate rank: this should cause rank improvements to be reduced by around 2 spots, but voters only worsen them by less than 1 spot. Voters do a better job of responding to social information for losing teams; the only evidence of significant underreaction occurs when top 15 teams have a better aggregate rank. We do not find any evidence that voters overreact to social information. The fact that voters are more Bayesian in response to losses is consistent with previous research showing individuals are more responsive to less ambiguous information (Sloman, Fernbach, and Hagmayer (2010), Rabin and Schrag (1999)), as losses are relatively salient and unambiguous signals (as compared to wins) for top 25 teams.

Because underreaction to social information is the predominant result, we introduce the term "stubbornness" to describe the voters' behavior. This term captures the idea that voters do not heed the information of others as much as they should. It can be thought of as describing a specific type of *conservatism*, a term commonly used in the belief updating literature to refer to underreaction to new information in general (Edwards (1968)).[6] However, we cannot conclude voters are stubborn purely due to information processing limitations, as voter behavior is likely also affected by reputation concerns. The theoretical effects of reputation concerns on responses to social information are ambiguous, as we discuss in section 2; when individuals care about reputation they may want to blend in with the crowd, or stand out from it, depending on the context.

We do not have a silver bullet for identifying reputation effects separately from information effects. However, we do have a few pieces of suggestive evidence, indicating that reputation

[6] We are unaware of an existing term in the literature equivalent to stubbornness. This may be because the behavior stubbornness–learning from others, but less than the Bayesian amount–is one that is rarely studied.

concerns cause “conformity”, or excessive similarity to others.[7] First, when we compare voter reactions to social information to their reactions to other types of new information, such as whether a game occurred at home or on the road, which should be less affected by reputation concerns, we find voters are more responsive to social information. Second, voters with very little poll experience, and thus relatively weak and uncertain reputations (causing strong reputation motives), are more responsive to social information. Third, voters are more responsive to social information in the later seasons of the sample, when analysis and criticism of individual voter rankings on the Internet was more intense, making reputation concerns likely stronger.[8]

Regardless of the composition of reputation and information processing effects, our main result, which seems quite robust, is that voters do learn from their peers, but less so than the Bayesian amount. We briefly discuss external validity and implications for future research in our concluding remarks.

2 Related Literature

There are several literatures that are especially relevant to this paper, in particular, the theory literatures on herding and reputation, and the empirical and experimental literatures on information processing and social learning. Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992) are seminal theory papers. They both show that a sequence of rational decision-makers often herds, i.e., individuals copy the decisions of those who acted before them, due to social learning. As discussed in the introduction, there is substantial empirical evidence that individuals do learn by observing the actions of others, but there is a lack of evidence on whether this learning is Bayesian. The experimental literature on social learning addresses this issue more directly than other literatures, but as far as we are aware results tend to be highly context-dependent. In some studies the Bayesian model seems to be a good description of behavior, while in others subjects imitate others excessively (e.g., Anderson and Holt (1997) and Offerman and Schotter (2009)), while in others still, subjects ignore their peers excessively (e.g., Çelen and Kariv (2005)). Similarly, it is still unclear to what extent individuals are ratio-

[7] The term conformity is used in different ways in existing literature. The term is generally used to refer to individuals acting to comply with a social norm or fit in with others; e.g., Bernheim (1994) or Corazzini and Greiner (2007). These papers do not take a stand on whether conformity is good or bad, with respect to the conformist. We use the term to refer to the opposite of stubbornness, which is excess reaction to social information. This usage is consistent with the term’s vernacular usage, implying excess compliance.

[8] Of particular note is the website pollspeak.com, which was founded in 2007 and calls itself “a watchdog organization dedicated to keeping college sports polls (and computer-rankings) honest”.

nal belief updaters to new information in general. Psychological theories and the experimental evidence are somewhat mixed. Salience and vividness seem to make individuals more responsive to new information and use, e.g., the *availability* or *representativeness* heuristics, both of which tend to cause overreaction. There is also substantial evidence of individuals being insufficiently responsive to new information; this is attributed to, e.g., the *conservatism* and *anchoring* biases, and is more likely when information is more ambiguous. See Holt and Smith (2009) for a recent lab study. Finally, since these studies involve stylized settings, it's not clear how the results would generalize to real world applications.

The theory literature has shown rational individuals may be either more or less likely to imitate their peers' actions when they have career concerns and are motivated by reputation. For example, Zwiebel (1995) shows that managers may herd because doing so reduces relative risk; when a manager takes the same action as her peers, she ensures that she will neither do much better nor worse than the pack. This may be relevant to AP voter behavior–if the voters' main objective is to continue to serve on the poll rather than rank teams in a risky way, this may cause voters to avoid unconventional rankings. On the other hand, the anti-herding literature (Levy (2004)) predicts individuals will excessively ignore or even contradict public information to signal confidence in the accuracy of their private information. This theory would predict voters respond insufficiently to the aggregate rankings. The empirical literature in this area focuses on the behavior of financial analysts and generally supports herding rather than anti-herding (Hirshleifer and Teoh (2003)), but there is evidence that individuals exaggerate their differences from peers (Zitzewitz (2001)).

Both theory and evidence indicate that reputation-motivated imitation becomes less likely as individuals gain experience. A natural explanation is that as experience is gained, reputation becomes more secure, and so the marginal effect of actions on reputation declines. This implies actions should be less driven by reputation concerns for those who are more experienced. See also Avery and Chevalier (1999), who provide a slightly different rationale for the prediction that herding declines with experience. Hong, Kubik, and Solomon (2000) provides empirical evidence in support of this idea, showing that less experienced stock analysts are both less likely to deviate from consensus forecasts, and more likely to lose their jobs if they do deviate.

There are several other strands of research on psycho-social reasons individuals may imitate each other more than the Bayesian amount, including the desire to attain status or esteem, social pressure, or simply because they gain utility from being similar to others (Bernheim (1994), Garicano, Palacios-Huerta, and Prendergast (2005), Zafar (2009)). We think that these are not likely to be major factors in our context for several reasons.[9] However, we note that if we do find evidence of conformity, we cannot rule these factors out as possible explanations. Finally, this paper relates to the other economics studies that also use college football rankings data, see, e.g., Mirabile and Witte (2009), Coleman, Gallo, Mason, and Steagall (2009) and Paul, Weinbach, and Coate (2007). None of them but none of them focus on social learning.

3 Empirical Strategy

Since our empirical method is an adaptation of Stone (2009), we have only included the most important points here, and refer the reader to that paper for additional detail.

3.1 The Data

The AP college football poll is conducted once per week during the college football season and teams usually play one game per week and sometimes have the week off. The season runs 16 or 17 weeks, and the first poll is conducted before the season starts, and the final poll after the season ends. The poll is voted on by 60-65 leading college football journalists (the number varies year to year), most of whom work for newspapers throughout the U.S. A small percentage work for television stations and other forms of media.

Each poll member votes by submitting 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). During the season, the current week's aggregate rankings are published in most newspapers and the current and historical aggregate rankings are available on many websites.[10] Consequently, prior to updating rankings each week, each voter is able to easily observe the Borda sum of her/his peers' rankings from the current and previous weeks.[11]

[9] For example, the AP voters do not interact with each other in person on a regular basis, and these factors are more likely to be relevant in situations involving personal interactions.

[10] See, e.g., appollarchive.com.

[11] The vast majority of games are played on Saturdays, throughout the day and evening, and the rankings are submitted to the AP the following day by 11:00 AM EST. Voters could consult with one another before submitting their updated rankings, but they would only have time to have extensive discussions with at most a few other voters, so we think it is safe to assume that the aggregate rankings potentially reflect a great deal of new information for all voters. The data indeed bear this assumption out, as we show the aggregate information has substantial effects on estimated Bayesian ranking responses.

The individual voter rankings are not confidential, but not widely available. Most of the data used for this study are the same as those used by Stone (2009), which describes how the data were obtained.[12] We obtained additional data on voter experience by directly communicating with the voters via email.

One notable weakness of the data is that the voters do not have direct incentives relating to the quality of their rankings. We do not believe this is too serious a concern for several reasons. First, discussions with voters indicate that they put substantial effort into producing their best possible rankings. Second, voters do have indirect career-related incentives based on the quality of their rankings. It is considered prestigious to be part of the poll, and voters are invited to be on the poll for multiple seasons in part based on their performances in previous polls.[13] In addition, voters' weekly ballots are scrutinized increasingly carefully by bloggers and websites, giving voters stronger incentives to be thoughtful about their rankings to avoid criticism.

3.2 The True Rankings and Identification

In order to identify systematic differences, or lack of difference, between observed voter responses and the Bayesian responses to social information, we need to first estimate the Bayesian responses to social information. In order to do this, we need a measure of the true rankings. We assume the true rankings for each voter-season are her or his final rankings for that season.

We need to make this assumption because the true rankings are not clearly defined.[14] The logic behind the assumption can be described informally as follows. We think of the voters as taking their best guess at what their own final rankings will turn out to be when they submit their weekly rankings. The voters are doing this not necessarily to forecast the future of the season. Rather, the voters are effectively guessing their final rankings because they know their own final rankings are the most accurate rankings that will ever be available, by the voters' own standards, for that season. This is because the final rankings reflect the maximal amount of information that will ever be available on the qualities of the teams that season, the qualities

[12] The only website that maintains an archive of historical individual voter rankings that we are aware of is pollspeak.com.

[13] An extreme example of a voter's performance affecting his participation in the poll occurred in 2006, when a voter was removed from the poll in mid-season after mistaking a win for a loss (http://sports.espn.go.com/ncf/news/story?id=2663882).

[14] Montella said in a discussion in the summer of 2009 that the rankings criteria are ambiguous. The voters are given the following guidelines, intentionally left open to interpretation, before each season: "Base your vote on performance, not reputation or preseason speculation. Avoid regional bias, for or against. Your local team does not deserve any special handling when it comes to your ballot. Pay attention to head-to-head results. Don't hesitate to make significant changes in your ballot from week to week. There's no rule against jumping the 16th-ranked team over the eighth-ranked team, if No. 16 is coming off a big victory and No. 8 just lost 52-6 to a so-so team." The first sentence was added for the 2008 season, and Montella said it was not indicative of a policy change, but just meant to encourage the voters to be responsive to game results.

of their performances, and the voters' own possibly idiosyncratic tastes in evaluating these qualities. It is worth noting that regardless of the assumption's validity, our estimates of voter over/underreaction (sections 4.2 through 4.5) can be interpreted as deviations from the goal of matching a voter's own final ranks for that season.

A natural question to ask is if the voters are, or should be, trying to guess their final rankings each week, why does the AP not instruct them to do this explicitly. The answer is that this would give the voters greater incentives to vote strategically–to adjust their final rankings so they are similar to their mid-season rankings, to make their mid-season rankings look more accurate. This would raise suspicion and reduce the legitimacy of the rankings in general. When it is unspoken that the final rankings are proxies for true rankings, these strategic issues are much less severe.

We briefly and informally discuss alternative possibilities for the true rankings. One plausible alternative criterion for the rankings is year-to-date (YTD) performance. This would be, after week 1, just week 1's performance, after week 2, week 1 and 2 performance, etc. If this were the criterion, the true rankings would change throughout the season. One reason it is unlikely the voters actually use this criterion is simply that there exists a preseason poll. Since this poll could not be based on YTD performance, it would be inconsistent for other polls to be based entirely on YTD performance. Another reason it is unlikely YTD performance is the only criterion is that if it were, voters would change their rankings drastically after games in the early weeks of the season. Simple inspection reveals this is not the case.[15]

It is possible the true rankings change throughout the season if voters rank teams based on current quality, and quality changes substantially throughout the season. This would imply the final rankings are not the true rankings for each week of the season. Stone (2009) provides evidence against this possibility. The paper shows that historically, teams that finish ranked 1-12 do not improve during the season more than teams that finish ranked 13-25. (If team qualities did change throughout the season substantially, and voters ranked teams on current quality, teams that finish ranked 1-12 should have shown more within-season improvement than lower ranked teams.)

[15] For example, Clemson lost its first game of the 2008 season by 24 points, but still received 143 points in the subsequent aggregate poll, more than that of over 90 other teams, most of which had won their first game or lost by fewer points.

Another plausible alternative definition of the true rankings is the aggregate final rankings. Given the premise of this paper–that individuals have heterogeneous information, and thus may learn from each other–the aggregate final rankings may be thought to be the best estimate of the true rankings by a law of large numbers-type argument. It is also possible that voters have incentives to conform to the aggregate rankings for reputational reasons, as discussed in section 2. If the voters considered the aggregate final rankings the true rankings, then clearly social information would be *more* important than if the voters simply considered their individual final rankings to be the true rankings. Thus, by assuming that individual final rankings are the true rankings in our analysis, our results should be biased towards finding *overreaction* to social information. Since, as discussed in section 1, our main finding is underreaction, we think we are on safe ground. We also discuss robustness to this issue in section 4.5.

We note that defining truth as the voters' own final rankings implies truth is endogenous: each voter determines his/her own truth. To address this, we restrict the sample we use for hypothesis testing to the first half of the seasons (weeks 1-7). This causes truth to be exogenous for practical purposes, as the rankings change considerably over the season's second half. We also note that this issue only potentially biases our analysis *against* rejecting the null of Bayesian updating, and so it does not seem to threaten the validity of our evidence against non-Bayesian social learning.

3.3 Formal Framework

In this subsection we specify the voters' objective functions and Bayesian updating process in a way that allows us to estimate the benchmark Bayesian posterior rankings that conditions on aggregate rank as a measure of social information.

Let r_i denote the true rank of team i (in a particular season and for a particular voter; those indexes are suppressed), $i \in \{1, ..., N\}$ and $r_i \in \{1, ..., N\}$, with N denoting the total number of teams. Let $\widetilde{r}_{i,t}$ be the actual (observed) rank that the voter assigns to team i in week t, in which $\widetilde{r}_{i,t} \in \{1, ..., 25, unranked\}$, since the voters can only rank 25 teams. We assume each voter's objective in week t, for all $t \in \{1, ..., T-1\}$ with T denoting the number of weeks in the season, is to minimize the expectation of a standard quadratic loss function of current and true ranks: $E_t[\sum\limits_{i=1:N} (\widetilde{r}_{i,t} - r_i)^2]$, with $\widetilde{r}_{i,t}$ equal to any number greater than 25 if in fact

$\widetilde{r}_{i,t} = unranked$. It can be shown this loss function is minimized only if teams are ranked in order of expected rank, with teams that do not have one of the 25 best expected ranks simply unranked. It is assumed voters are not forward-looking (they do not consider adjusting future ranks to reduce previous losses).

Voters update their beliefs about the true rankings throughout the season by observing the aggregate rankings and game scores. Let a_t be the vector of aggregate rankings in week t and s_{ij} be the score difference for the game in which home team i plays team j: the points scored by i minus points by j (if $s_{ij} > 0$, i wins).[16] In accordance with the timing of the rankings submissions discussed above, we can think of each week as starting Sunday evening and ending the following Sunday morning. Voters submit their week $t+1$ rankings at the very end of week t, and the week $t+1$ rankings of all voters are aggregated and made public at the very start of week $t+1$. Consequently, in each week t, voters observe a_t at the very start of the week. Since games are Saturdays, voters observe s_t, the vector of scores from week t, towards the end of the week and just before submitting their week $t+1$ rankings. Let I_t denote information available at the start of week t, when the aggregate rankings for that week are made public. The timing of the set-up implies $a_t \in I_t$ and $s_t \notin I_t$, and that the week t rankings are conditioned on all information from week $t-1$ (I_{t-1} and s_{t-1}).

Let $f_i(r_i|I_t)$ be the (subjective) probability that team i has true rank r_i conditional on information set I_t. $f_i(r_i|I_t)$ is the week t "prior" for i in that it is the probability prior to the game result signals of week t. Let $g(s_{ij}|r_i, r_j)$ denote the conditional probability that the game between teams with true ranks r_i and r_j results in score s_{ij}, with i being the home team. Assuming voters know the $f()$ and $g()$ distributions from their knowledge of the sport, they can use Bayes' rule to update beliefs (their subjective true rank probabilities) after team i plays j and s_{ij} is observed, from $f_i(r_i|I_t)$ to $f_i(r_i|s_{ij}, I_t)$.

Specifically, assuming for simplicity $f_{ij}(r_i, r_j|I_t) = f_j(r_j|I_t)f_i(r_i|I_t)$, the formula for a Bayesian posterior is:

$$f_i(r_i|s_{ij}, I_t) = \frac{g(s_{ij}|r_i, I_t)f_i(r_i|I_t)}{g(s_{ij}|I_t)} = \frac{\left[\sum\limits_{r_j} g(s_{ij}|r_i, r_j)f_j(r_j|I_t)\right] f_i(r_i|I_t)}{\sum\limits_{r_i}\left[\sum\limits_{r_j} g(s_{ij}|r_i, r_j)f_j(r_j|I_t)\right] f_i(r_i|I_t)}. \quad (1)$$

[16] This variable has no time subscript because teams almost never play each other more than once.

We can use this formula to estimate the voters' Bayesian posterior beliefs, given the availability of estimates of the prior and signal distributions. We can then translate these estimated posterior beliefs into rankings, by ordering them by posterior expectation given the quadratic loss function assumption, to obtain the estimated Bayesian posterior rankings.

One aspect of this extended framework worth noting is that we are assuming voters actually update beliefs in two stages in each week t–once at the start of the week after observing a_t, and once at the end after observing s_t. However, we only explicitly apply the Bayesian updating formula to estimate the voters' posterior beliefs once, to update beliefs based on scores. The priors used for this estimation are conditional on a_t, which makes them implicitly Bayesian. This is without loss of generality and makes the computation simpler.

3.4 Estimation Methodology

We use empirical frequencies to estimate the f's and g's in order to apply (1) to obtain the Bayesian posteriors. To estimate the f's–the true rank distributions–we use the final rank frequencies, separately for each current rank, week and relative aggregate rank. To estimate the g's–the conditional score distributions–we use the frequencies with which opposing teams with various final ranks had different scores. See the supplementary appendix for details. A few adjustments forced by data and sample size limitations that should be noted are the following. We coarsen some of the variables and only estimate posterior rankings for teams that are currently in each voter's top 25 (there are over 100 teams).[17] Also, the score distributions are conditioned on aggregate final rank and the 2006-08 data are used to estimate the priors. The former assumption could cause our estimated posteriors to be invalid (biased in a way that is theoretically unclear); we show evidence that this is not a problem in section 4.1. We test robustness to the second issue (using 06–08 data) by also conducting the analysis using priors estimated from the 2006-07 data only on a sample of 2008 only data. We get similar results, discussed in section 4.5, though they are much less precise due to the lost sample.

The main modification of the existing framework is that the information set the estimated priors are conditioned on is expanded to account for aggregate rank. As discussed above,

[17] We coarsen final rank into four groups, 1-6, 7-12, 13-18 and 19-25, and scores into bins of width seven. To illustrate how the $f()'s$ are estimated by example, the estimated probability a team with current rank 25 in week 1 with $AggW = 1$ (defined below) has true rank 1-6 ($f_{1-6}|I_1$), we simply use the frequency (fraction) of the teams ranked 25 in week 1 with $AggW = 1$ that indeed finished the season with rank 1-6 by the same voter who originally ranked the team 25. We account for the fact that teams enter/exit voters' rankings in most weeks by only ranking teams up to the number of teams that stay in each voter's rankings, for each week. We use the same method for comparing estimated to actual final rankings in section 4.1.

voters can easily observe the aggregate rankings for the current week before submitting their rankings for the following week. We account for aggregate rank in a simple way, adding two dummy variables to I_t: *AggB* and *AggW*. $AggB = 1$ implies the aggregate rank is "better" than the voter's own rank for that team, $AggW = 1$ implies the team's aggregate rank is "worse". When $AggW = AggB = 0$, the aggregate rank is the "same". Since there is no clear theoretical framework for defining better, worse and same, we experiment with a few *ad-hoc* but straightforward definitions. Naturally *AggB* (*AggW*) equals one only if the aggregate rank is at least one rank better (worse) than the individual rank. We conducted our analysis with numerous definitions, and report results for four that are especially straightforward: *AggB* (*AggW*) equal 1 if aggregate rank is at least two, three, four or five spots better (worse) than the voter's individual rank.[18] The trade-off among definitions is that the greater the difference between the individual rank and the aggregate rank, the more informative it is for *AggB* or *AggW* to equal one, but the less informative it is when $AggB{=}AggW = 0$. Moreover, conditioning on a larger difference between the aggregate rank and the individual rank means that we have fewer observations in the sample to estimate priors for teams with *AggB* and *AggW* equal to one, making the estimates less precise.

4 Empirical Analysis

4.1 Validity of the Estimated Bayesian Posteriors

Before testing whether the estimated Bayesian posteriors (henceforth the *Bayesian posteriors* or *estimated posteriors*) are systematically different from the observed voter posterior rankings (the *observed posteriors*), we first examine the validity of the Bayesian posteriors. If our estimates are not valid–if they are not reasonably unbiased and precise measures of Bayesian reactions to social information–it would be difficult to use the estimates to draw conclusions on whether or not the voters are Bayesian social learners.

We examine validity by comparing the distances between the Bayesian posteriors and the final rankings, to the analogous distance between the observed posteriors and final rankings. The true Bayesian posterior rankings, which are of course unobserved, will on average be closer to the true rankings, as compared to the distance between any other posterior rankings and

[18] We also used definitions in which the cut-offs varied depending on the team's rank; we found these yielded similar results so we only report results for the most straightforward definitions. Results for other definitions are available upon request.

the true rankings. Thus, since we assume the final rankings are the true rankings, if our estimated posteriors "match" the final rankings better than the observed posteriors, this would be evidence of the validity of the estimates. It is worth noting that this is quite a strong condition for validity, since it requires both that the estimates use the information they are conditioned on in a more rational way than the voters use that same information, and that the information available to the voters, but not incorporated in the estimates, is not too important to outweigh the gains from the improvement in rationality.

We also compare the distances from truth for the "non-social" Bayesian posteriors (estimated posteriors that are only conditioned on game scores and not aggregate rankings), observed priors and flat priors. It is important to see whether the Bayesian posteriors that are conditioned on social information are closer to the final rankings than the non-social posteriors, to determine whether the aggregate rankings are actually informative with respect to the voters updating towards their individual final rankings. If the aggregate ranks were uninformative, the social and non-social Bayesian posteriors would be equally close in distance to the true rankings. We include the observed and flat priors as benchmarks to see how informative game scores and the priors are in general.

We measure the distance from the final rankings by mean absolute deviation (MAD): $\frac{1}{n}\Sigma_{i,t,s,v}|\widehat{r}^{v}_{i,t+1,s} - r^{v}_{i,s}|$, with n being the total number of observations, and i, t, s and v indexing team, week, season and voter, respectively. $\widehat{r}$ is the estimated or observed posterior, or prior; $r^{v}_{i,s}$ is voter v's true rank for team i in season s, which is not indexed by t as true rank is constant by season. We could use other similar metrics and obtain similar results; this method is less susceptible to influence by outliers.

Table 1 presents the MADs, split out by game result type and rank group. In general–for almost all game types and rank groups–the Bayesian posteriors are closer to the final rankings than the observed posteriors. The overall ("Total") MADs for Bayesian posteriors range from 3.33 to 3.40 for the four $AggB/AggW$ definitions; for the observed posteriors the overall MAD is 3.68. Moreover, the Bayesian posteriors that are conditional on social information are always closer to the true rankings than the non-social Bayesian posteriors (overall MAD of 3.49). This implies the aggregate rankings are informative for the voters updating towards their subjective true rankings. This is a non-trivial result–it is not, say, analogous to increasing a multiple regression R^2 by adding variables–as conditioning on more information is costly in that it decreases the sample sizes for prior estimation. While these comparisons are informal, the fact that the MADs of Bayesian estimates conditional on social information are consistently and substantially lowest is strong evidence of the validity of the Bayesian posteriors.[19]

The MADs also provide a criterion for selecting the preferred *AggB/AggW* definitions to use for the subsequent formal testing analysis. We wish to use the definitions that yield the "most Bayesian" posteriors. This is the definition that yields lowest MADs on average. Table 1 shows that, according to this criterion, the first two definitions (henceforth, Definitions 1 and 2) are superior. In the remainder of the paper, we only report results for these two definitions; results using other definitions are qualitatively similar.

Table 2 provides a preview of the formal analysis, reporting summary statistics on rank improvement, categorized by the signal type (wins; losses) and broken out by prior rank categories, and by *AggB*, *AggW*, and $AggS = 1 - AggB - AggW$. The mean estimated and observed rank changes are similar in the overall sample. However, in the overall sample, the observed rank improvements are smaller (larger) in the case of wins (losses) when *AggB* equals one, as compared to the estimated Bayesian improvements. There are also a few notable discrepancies within some of the specific rank groups. For losing teams ranked in the top 15 with *AggB*=1, the estimated and observed responses are very different: the observed rank decline is much larger than the estimated Bayesian rank decline. For winning top 15 teams with *AggW*=1, observed rank decline is smaller than the estimated decline. In the case of winning teams with prior rank 16-25, observed rank improvements are smaller (larger) than the estimated rank improvements when *AggB* (*AggW*) equals 1. All of these numbers suggest a pattern of voters not incorporating social information in their ranking updates as much as they should.

4.2 Hypothesis Testing

To conduct formal hypothesis tests, we first construct a simple measure of overreaction, denoted *Over*, intended to measure excess rank improvement (decline) following a win (loss). It is defined

[19] We do not conduct formal tests of these differences because the observations are correlated across voters for games involving the same teams. If we did not account for this correlation, the differences would easily all be significant due to the large sample sizes. We could conduct the tests of differences separately for each voter to account for this issue, but this would make the presentation of results even more unwieldy. The Bayesian posterior MADs for all *AggB/AggW* definitions are significantly lower than the observed posteriors for a substantial percentage of voters.

as follows:

$$Over_{it} = \begin{cases} \Delta r_{it}^{O} - \Delta r_{it}^{B} = r_{it}^{B} - r_{it}^{O} \text{ if team } i \text{ wins in week } t, \\ \Delta r_{it}^{B} - \Delta r_{it}^{O} = r_{it}^{O} - r_{it}^{B} \text{ if } i \text{ loses in } t, \end{cases}$$

in which $\Delta r_{it}^{j} = r_{it-1} - r_{it}^{j}$ ($j \in \{O, B\}$, $O =$ observed; $B =$ Bayesian) is the rank improvement for team i in week t.[20]

We then use *Over* to test whether voters are Bayesian social learners by estimating linear models of the form:

$$Over_{ivts} = \beta_1 AggB_{ivts} + \beta_2 AggW_{ivts} + X_{ivts}\lambda + \delta_v + \gamma_s + \eta_t + \epsilon_{ivts}. \tag{2}$$

δ_v, γ_s, and η_t are voter, season and week fixed effects (FEs), respectively. The voter FEs are used as controls to account for voter-specific variation in priors causing them to appear to over or underreact. We use bootstrap standard errors since the dependent variable is constructed, and the standard errors are clustered by game to account for repetition of games in the sample and correlation in the error term by game. X is a vector of controls, and includes the following in all of our specifications:

1) *Home*: dummy for team i playing a home game in week-season t-s;

2) *ScoreMargin*: team i-t-s's points minus its opponent's points;

3) *OppRank*: dummy for whether i-t-s's opponent is ranked in v's top 25 in t-s;

4) *ScoreMargin* × *OppRank*.

The parameters of interest are β_1 and β_2. The null hypothesis is that voters are Bayesian social learners and that both of these parameters equal zero. The alternatives are the following.

Alternative Hypotheses to Bayesian Social Learning

	Wins	Losses
Conformity	$\beta_1 > 0$, $\beta_2 < 0$	$\beta_1 < 0, \beta_2 > 0$
Stubbornness	$\beta_1 < 0, \beta_2 > 0$	$\beta_1 > 0$, $\beta_2 < 0$

Conformity and stubbornness are blanket terms we use to refer to overreaction and underreaction to the aggregate rankings, respectively.[21] Hirshleifer and Teoh (2003) define a taxonomy in which the term "herding" refers to individuals taking similar actions as their peers, and "dis-

[20] Note that all r variables are adjusted to account for the number of teams which enter/exist the top 25 for each voter in each week; see section 3.4.

[21] They are blanket terms in that they describe broad types of behavior; neither term is meant to describe the cause of behavior.

persing" refers to individuals taking opposite actions from their peers. We use herding with the same meaning, and anti-herding as they use dispersing, but these terms are not adequate for describing our hypotheses. This is because we are not interested in the relatively simple question of whether voters imitate each other or not; we are interested in whether voters herd excessively, or insufficiently, with respect to a Bayesian benchmark.

We say voters conform when they herd excessively.[22] We say that voters are stubborn when they do the opposite–herd insufficiently. We are agnostic *a priori* as to whether to expect conformity or stubbornness. Both may result from either non-Bayesian belief updating and/or reputational motives, as discussed in section 2.

Our main objective is simply to identify whether voters are Bayesian social learners on net, and if not, in what direction they deviate. We cannot cleanly distinguish information processing from other factors that may influence voter behavior. However, we can shed some light on the importance of reputation concerns in two ways. First, we compare the effects of aggregate rank (social information) to the effects of other types of new information, such as score and home status, on voter ranking updates. We do this to exploit the natural assumption that reputation concerns would have a relatively large effect on voter responses to social information, as compared to their effect on voter responses to other information. If we find voters are indeed relatively highly responsive to social information, this would be evidence that reputation motives cause herding (even if the net response to social information is less than the Bayesian amount). If voters respond even less to social information than to other information, this would be evidence that reputation causes anti-herding.

We cannot simply compare the estimates of β_1 and β_2 to our other estimates (λ) for this purpose, because they have different units. We create a unitless metric for these comparisons by estimating auxiliary regressions of the same form as (equation 2), but using the Bayesian rank improvement (decline) $\Delta r^B_{i,t}$ (-$\Delta r^B_{i,t}$) in the case of wins (losses) as the dependent variable, and then dividing our original coefficient estimates (from (2)) by the auxiliary estimates. The coefficient estimates from the auxiliary regressions are estimated effects on Bayesian rank change. The estimates of (2) are estimated effects on the difference between Bayesian and observed rank change. Thus, the ratio of estimates for a particular variable provides a unitless measure of the

[22] This occurs when they improve rankings too much after wins by teams with better aggregate rankings, improve rankings insufficiently after wins by teams with worse aggregate rankings, and worsen rankings excessively (insufficiently) after losses by teams with worse (better) aggregate ranks.

degree of non-Bayesian reaction, to that variable (a ratio of zero implies Bayesian updating). We refer to the absolute values of these ratios as *reaction ratios.*[23] For example, if for the wins sample we estimated β_1 to be -0.6 and 1.2 for the original and auxiliary regressions, the reaction ratio would be 0.5. If we obtained estimates of λ_{Home} of 0.9 and 1, this would yield a reaction ratio of 0.9.[24] These results would suggest that while voters do not respond to $AggB$ the full Bayesian amount, as the -0.6 estimate implies voters improve $AggB = 1$ teams' ranks by 0.6 spots less than they should after wins, voters respond to social information more strongly than they do to home status, since 0.5 is closer to zero than 0.9 is. To be clear, we fully acknowledge that these comparisons are not apples-to-apples and are merely suggestive, and we are cautious with the conclusions we draw from them.[25]

We can also gain insight into the effects of reputation by estimating the effects of several interaction terms. The effects of voter experience on the effects of $AggB$ and $AggW$ are of particular interest. In the absence of reputation effects, we would expect voters to become more Bayesian with experience simply due to improvement in understanding the prior and signal probabilities, which would imply a decrease in the magnitudes of the estimates for β_1 and β_2. But as discussed in section 2, both theory and evidence suggest that reputation-motivated herding declines with experience. Thus, if we find voters are stubborn (conformist) in general, and weakly more so as they gain experience, this would be suggestive that reputation concerns cause herding (anti-herding).[26] Other variables that may provide insight into the importance of reputation via interaction terms are year of season (as voter scrutiny on the Internet increased over the sample time-frame) and voter news organization affiliation (as reputation concerns may be larger for voters who work for national organizations).

4.3 Main Results

Table 3 reports estimates of several variants of equation (2), with reaction ratios reported in square brackets. Columns (1) and (4) report estimates of equation (2) for the wins sample for $AggB/AggW$ definitions 1 and 2, respectively. The estimates of β_1 are negative and β_2 are

[23] We use absolute value because we are interested in the relative magnitude of overreaction/underreation, as compared to Bayesian reaction. This comparison is only appropriate for two variables that the voters either (weakly) underreacted, or overreacted to.

[24] This implies underreaction to home status, since being the home team makes a win a less positive signal; see Stone (2009).

[25] We do not even formally test whether the ratios are significantly different from one another as estimating their standard errors would be difficult. We believe they provide useful suggestive evidence regardless.

[26] If voters became either less stubborn or conformist as they gained experience, we would not be able to say whether the change in behavior was due to a change in reputation concerns or improved (more Bayesian) information processing.

positive, and three of four are significant at either 5% or 1%. Columns (7) and (10) report the corresponding estimates for the losses sample. The β_1 estimate is significantly positive at the 5% level, while β_2 is negative but insignificant. According to the hypotheses described above, these estimates all indicate stubbornness, especially when $AggB = 1$.

To interpret the magnitudes of these coefficients, we look at the reaction ratios. For *AggB*, in the case of a win, the ratio is around 0.5; this means that voters respond only half as much to *AggB* relative to what Bayesian updating would have implied. Since the reported coefficient on *AggB* is around -1.78, the effect *AggB* has on Bayesian improvement must be around 3.5. The ratios for *AggW* for the wins models are 0.83 and 0.74. The ratios for *Home* and *Score Margin* variables are around 1 and 0.69, respectively.[27] Since the magnitudes of these ratios are generally greater than the ratios for *AggB*, and for *AggW*, this suggests that, although voters react to social information less than the Bayesian amount, they still react more strongly to social information than they do to other information.

Table 3 also reports results from specifications involving interactions of *AggB/AggW* with dummies for whether or not the opponent is ranked, and prior rank group. Games against ranked opponents receive much more attention as they occur relatively infrequently and are more informative regarding team qualities,[28] so it is possible voters pay more attention to these games and adjust their rankings differently, and perhaps in a more Bayesian way, in response to them. These results are reported in columns (2) and (5) for the wins sample. Voters underreact to *AggB* and *AggW* whether the opponent was ranked or not. The magnitude of the coefficient for *AggB* is larger for unranked opposition, while the magnitude of the coefficient for *AggW* is greater for ranked teams. This suggests that voters underreact to *AggB* more when the opponent is unranked, and underreact more to *AggW* when the opponent is ranked. However, these differences are not statistically significant. None of the estimates for the *AggB/AggW* interaction terms for ranked and unranked opponents are significant at 5% for the losses sample (columns (8) and (11)). Overall, these results suggest that voters pay similar attention to social information when teams play ranked opponents or unranked opponents.

The specifications with interactions of *AggB* and *AggW* with the rank subgroups (1-5, 6-15, and 16-25) allow us to dig somewhat deeper into the results, and in particular, check to see that

[27] We focus on the ratios for *Home* and *ScoreMargin* since they are consistently highly significant; *OppRank* and *OppRank* × *Score* tend to be insignificant.

[28] This can be seen in Tables 1 and 2 as the MADs are lower and rank changes are larger, after losses as compared to wins.

the results hold for the mid-ranked teams, which are least affected by the censored nature of the rankings data. The ranks of teams ranked 6-15 can both improve and worsen substantially, despite the fact that there are upper and lower bounds to observed ranks. The interaction results indeed confirm that the overall trends hold for teams ranked 6-15, as both *AggB* and *AggW* are significant at 1% for wins samples (columns (3) and (6)), with reaction ratios less than 0.7. Also, *AggB* is positive and significant at 1% for losses for these teams (columns (9) and (12)), again indicating stubbornness (the other losses estimates are insignificant). It is also worth noting the reaction ratios indicate that underreaction is smallest for teams with ranks 16-25, implying voters are more responsive to social information for relatively weak teams.

4.4 Voter Heterogeneity and Additional Interactions

To get a sense of heterogeneity across voters in responsiveness to social information, we estimate the model in equation (2) with *AggB* and *AggW* interacted with voter fixed effects. This gives us estimates of β_1 and β_2 for each voter. As expected, these estimates are not very precise but an F-test rejects the null that these coefficients are jointly zero at the 1% level, suggesting substantial heterogeneity. This can be seen in Figure 1, which shows a scatterplot of voter *AggB* and *AggW* for wins (losses) in the left (right) panel. In the case of wins, there's a weak negative relationship between *AggB* and *AggW*, suggesting that voters have a tendency to either underreact or overreact to social information in the case of wins irrespective of whether the social information is better or worse than the voter's rank. That, however, is not the case for losses.

We next examine if voter observables can explain this heterogeneity and provide insight into the importance of reputation concerns. Columns (1) and (7) of Table 4 present the estimates of equation (2) for wins involving our first (of two) *Experience* variable interacted with *AggB* and *AggW,* for Definition 1 and Definition 2, respectively. This variable is defined as the number of years the voter has voted on the AP polls since 1998; its sample mean is 2.66. Both interactions with *AggW* are positive and significant at the 5% level. There are no significant effects in the case of losses, as shown in Table 5. Columns (2) and (8) of both tables report the estimates for an alternative experience interaction term: years in sports journalism.[29] In our sample, this measure of experience varies from 3 to 42 years with a mean of 21.39 years. The magnitudes of

[29] This information was collected by emailing each of the voters. We received replies and thus have data for 82 of the 122 voters.

these estimates are small, and none of them are statistically significant at the 5% level. These results indicate that while voters become more stubborn as they gain poll experience, their reaction to social information does not change with journalism experience.

About 5% of the voters in our sample are affiliated with national news sources, such as ESPN and Sports Illustrated. These voters likely have stronger reputational concerns, and their rankings may be scrutinized by fans and other media more closely, so we test for systematic differences in their social learning. Columns (3) and (9) present the estimation of equation (2) with the inclusion of *Nat* (a dummy that equals one if the voter is affiliated with a national news source) interacted with $AggB$ and $AggW$ for the wins and losses samples, respectively. Separate coefficients are estimated for voters with national and local (defined as *1-Nat*) affiliations. There are no consistent differences between $Nat = 0$ and $Nat = 1$ voters, and the null that the coefficients are different for national and local voters cannot be rejected at the 10% level.

Most newspapers devote a large fraction of their sports coverage to local teams. We define a dummy *Follow* that equals one for the game-voter observation if the game includes a team that the voter's news source has a dedicated section for on its website; this is the case for 3.66% of the observations. Voters may have more precise priors on these teams, in which case they should be less responsive to social information. Estimates of the model that include this dummy are shown in columns (4) and (10). We do not find any systematic difference in response to social information; the null that the coefficients are different for games including a team that is closely followed and for games that do not include a team that is closely followed (i.e., *1-Follow*) cannot be rejected at the 10% level. As before, the estimates are overall consistent with underreaction to social information.

Our sample includes some voters who voted on the poll in more than one year. We next analyze how response to social information changes over time for this group of voters. For this purpose, we restrict our sample to those voters who appear in at least two years in our sample period 2006-2008. We are left with 48 of the 121 voters for this analysis. It should be pointed out that the average years of experience (defined as number of years the voter has voted on the AP polls) is higher for this subgroup relative to the entire sample. We define a dummy *Repeat* that equals one if the observation comes from the second or third year of the voter's participation in the polls. Estimates of the model that includes this dummy variable are reported in columns (5) and (11). The estimates show that voters' response is similar (underreaction) to $AggB$ irrespective of whether $Repeat$ equals 1 or 0. The interaction effects for $Repeat = 0/1$ are statistically indistinguishable from each other for the losses sample (columns (5) and (11) of Table 5), and for the wins sample using definition 1 (i.e., column (5) in Table 4). However, for the wins sample, Definition 2, the coefficients on $AggW$ are statistically different: relative to response in later years, voters underreact more to social information in their first year (during the sample period). This particular behavior is consistent with voters conforming more with experience. Since these voters on average have more poll experience than voters in the full sample, this indicates experience may have non-linear effects; voters may become more stubborn as they first gain experience, but less so after they have served on the poll a sufficient amount of time.

Finally, we examine year effects in columns (6) and (12). We control for poll experience, since this is also increasing with year for many of the voters, and we know it may have significant effects. For wins, the $AggW$ estimate is highest in 2006, and for losses, the $AggB$ estimate is highest then. Both of these results indicate voters are more responsive to social information in 2007-08 than 2006.

4.5 Robustness

We have shown our results are robust to using different definitions of $AggB/AggW$ and specifications of the regression model, but in this subsection we discuss two additional robustness checks. In the first, we only use the 2006 and 2007 seasons to estimate the priors, and only construct posteriors and conduct analysis using data from the 2008 season. This approach is preferable in that it allows us to check that aggregate information is not under-valued for just one particular season and prevents us from incorporating information in our estimates of priors not available to the voters; the downside to this approach is it reduces the sample sizes for all aspects of the estimation procedure. In the second robustness check, we use the aggregate final ranks, rather than individual final ranks, as the true rankings to estimate the priors. As discussed above, given that information is heterogeneous, it may be reasonable for voters to think that the aggregate rankings are more accurate than any individual's rankings, and thus to treat the aggregate final rankings as the best rankings for a given season. Table 6 reports the main results for these checks (coefficients on $AggB/AggW$ for basic regressions). It shows that results are very similar to those reported in Table 3, indicating robustness to these issues. The estimates are slightly smaller in magnitude for the first check, and slightly larger for the second, but they are directionally the same and still often significant at the 1% and 5% levels.

5 Conclusion

We have found strong evidence that AP college football poll voters are underreactive social learners. While the voters' peer rankings are informative with respect to individual voters determining their own final rankings, and the voters do adjust their individual rankings towards their peers' rankings to some extent, they do so less than the Bayesian amount. We refer to this behavior as stubbornness, since voters essentially are not listening to their peers as much as they should. Voters are more responsive to peer rankings of losing teams, as compared to winning teams. Since wins are more ambiguous (i.e., less informative) and less salient signals than losses, this behavior is consistent with existing research showing individuals are more likely to "stick to their guns" when it is easier to deny the contradictory evidence. Reputation concerns appear to cause voters to conform, though not enough to overcome the overall tendency to be stubborn, as voters react more strongly to social information than other information, and less experienced voters, and voters in 2007 and 2008 (when media scrutiny increased), respond more strongly to social information.

We hope this paper will enhance understanding of social learning, informing future theoretical work and policy discussion, along the lines of Cipriani and Guarino (2008), Golub and Jackson (2010) and Glaeser and Sunstein (2009). We note that there likely are limitations to the external validity of our findings. Our main result–that individuals do not listen to their peers as much as they should–may depend on a number of factors, perhaps in particular the fact that the subjects of our study are experts in their field. Individuals with less experience in an area may fail to be so confident, and so could be susceptible to the opposite mistake of being unduly persuaded by their peers. Furthermore, the incentives provided by reputation concerns may be relatively weak in the context we study, due to the lack of direct career-related incentives relating to quality of rankings. In other contexts with stronger reputation effects, individuals may conform even if they under-appreciate the information value of their peers' actions. Understanding the aspects of context that influence social learning is an important avenue for future research.

References

Anderson, L., and C. Holt (1997): "Information cascades in the laboratory," The American Economic Review, pp. 847–862.

Avery, C., and J. Chevalier (1999): "Herding over the career," Economics Letters, 63(3), 327–333.

Banerjee, A. (1992): "A simple model of herd behavior," The Quarterly Journal of Economics, 107(3), 797–817.

Bernheim, B. (1994): "A theory of conformity," The Journal of Political Economy, 102(5), 841–877.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992): "A theory of fads, fashion, custom, and cultural change as informational cascades," Journal of Political Economy, 100(5), 992.

Cai, H., Y. Chen, and H. Fang (2009): "Observational Learning: Evidence from a Randomized Natural Field Experiment," American Economic Review, 99(3), 864–882.

Çelen, B., and S. Kariv (2005): "An experimental test of observational learning under imperfect information," Economic Theory, 26(3), 677–699.

Cipriani, M., and A. Guarino (2008): "Herd behavior and contagion in financial markets," BE Journal of Theoretical Economics, 8(1).

Cohen-Cole, E., and B. Duygan-Bump (2008): "Household Bankruptcy Decision: the role of social stigma vs. information sharing," Federal Reserve Bank of Boston, mimeo.

Coleman, B., A. Gallo, P. Mason, and J. Steagall (2009): "Voter Bias in the Associated Press College Football Poll," Journal of Sports Economics.

Conley, T., and C. Udry (2010): "Learning About a New Technology: Pineapple in Ghana," The American Economic Review, 100(1), 35–69.

Corazzini, L., and B. Greiner (2007): "Herding, social preferences and (non-) conformity," Economics Letters, 97(1), 74–80.

Edwards, W. (1968): "Conservatism in human information processing," Formal Representation of Human Judgement, New York, John Wiley and Sons.

Garicano, L., I. Palacios-Huerta, and C. Prendergast (2005): "Favoritism under social pressure," Review of Economics and Statistics, 87(2), 208–216.

Glaeser, E., and C. Sunstein (2009): "Extremism and social learning," J. Legal Analysis, 1, 263.

Golub, B., and M. Jackson (2010): "Nave Learning in Social Networks and the Wisdom of Crowds," American Economic Journal: Microeconomics, 2(1), 112–149.

Hirshleifer, D., and S. Teoh (2003): "Herd behaviour and cascading in capital markets: A review and synthesis," European Financial Management, 9, 25–66.

Holt, C., and A. Smith (2009): "An update on Bayesian updating," Journal of Economic Behavior and Organization, 69(2), 125–134.

Hong, H., J. Kubik, and A. Solomon (2000): "Security analysts' career concerns and herding of earnings forecasts," The Rand journal of economics, 31(1), 121–144.

Knight, B., and N. Schiff (2007): "Momentum and social learning in presidential primaries," NBER WORKING PAPER SERIES, 13637.

Levy, G. (2004): "Anti-herding and strategic consultation," European Economic Review, 48(3), 503–525.

Mirabile, M., and M. Witte (2009): "Not So Fast, My Friend: Biases in College Football Polls," Journal of Sports Economics.

Moretti, E. (2008): "Social learning and peer effects in consumption: Evidence from movie sales," NBER Working Paper.

Offerman, T., and A. Schotter (2009): "Imitation and luck: an experimental study on social sampling," Games and Economic Behavior, 65(2), 461–502.

Paul, R., A. Weinbach, and P. Coate (2007): "Expectations and voting in the NCAA football polls: The wisdom of point spread markets," Journal of Sports Economics, 8(4), 412.

Rabin, M., and J. Schrag (1999): "First Impressions Matter: A Model of Confirmatory Bias*," Quarterly Journal of Economics, 114(1), 37–82.

Sloman, S., P. Fernbach, and Y. Hagmayer (2010): "Self-deception requires vagueness," Cognition.

Stone, D. (2009): "Testing Bayesian Updating with the AP Top 25," SSRN Working Paper.

Zafar, B. (2009): "An Experimental Investigation of Why Individuals Conform," Working Paper.

Zitzewitz, E. (2001): "Measuring herding and exaggeration by equity analysts and other opinion sellers," Working Paper.

Zwiebel, J. (1995): "Corporate conservatism and relative compensation," The Journal of Political Economy, 103(1), 1–25.

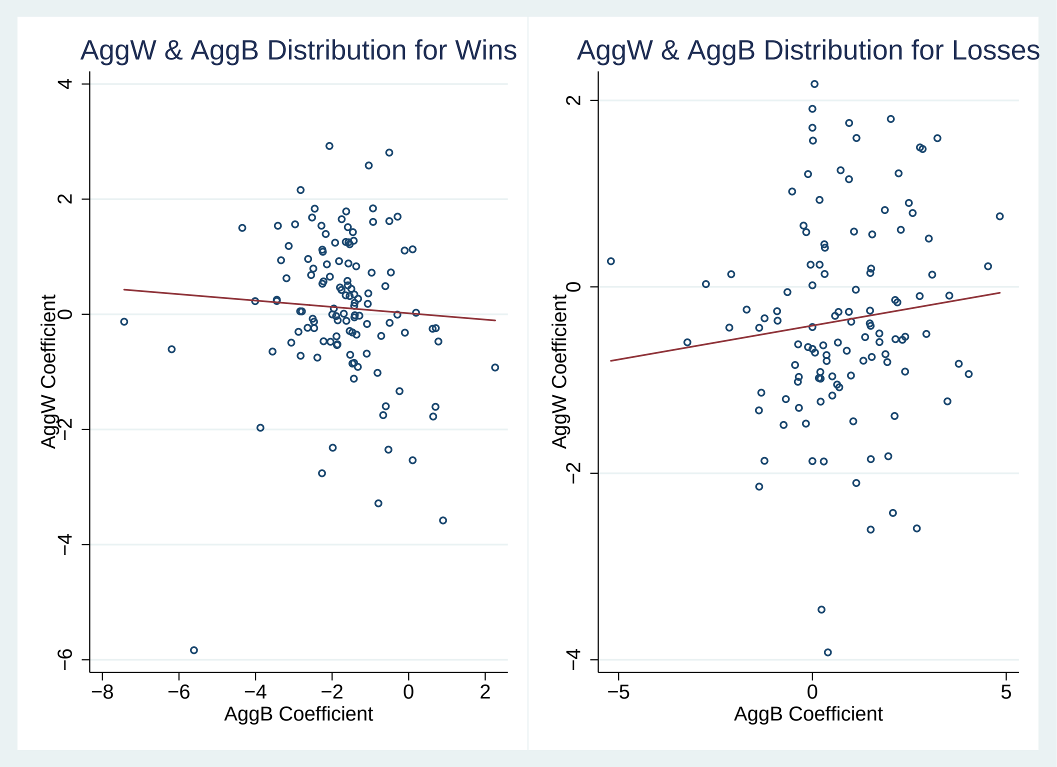

Figure 1: Distribution of individual voter *AggB* and *AggW* estimates for wins (left figure) and losses (right figure).

Table 1: Mean Absolute Deviations (MADs) from Final Ranks (Standard Deviations in Parentheses)

	Number of	|Cond. Est. Bayes. Post - Obs Final|				|Non-Social	|Obs Post -	|Obs Prior -	|Flat Prior
	Observations	Defn 1	Defn 2	Defn 3	Defn 4	Post - Obs Final|	Final|	Final|	- Final|
All Teams									
Wins	21,811	3.69	3.73	3.76	3.79	3.90	4.03	3.89	4.88
		(3.33)	(3.34)	(3.39)	(3.40)	(3.48)	(3.63)	(3.42)	(3.43)
Losses	6,680	2.07	2.10	2.12	2.12	2.13	2.38	3.27	3.75
		(3.14)	(3.24)	(3.26)	(3.27)	(3.28)	(3.40)	(3.24)	(2.80)
Byes	2,887	3.30	3.38	3.41	3.42	3.51	4.05	3.99	4.35
		(3.57)	(3.64)	(3.62)	(3.62)	(3.72)	(3.89)	(3.67)	(3.30)
Total	31,378	3.31	3.35	3.38	3.40	3.49	3.68	3.77	4.59
		(3.37)	(3.41)	(3.45)	(3.46)	(3.53)	(3.67)	(3.42)	(3.33)
Rank 1-5									
Wins	4,609	3.59	3.67	3.73	3.72	3.80	3.78	3.75	7.30
		(3.02)	(3.12)	(3.18)	(3.22)	(3.29)	(3.33)	(3.28)	(3.52)
Losses	1,605	3.85	3.92	3.89	3.82	3.81	4.88	5.83	5.89
		(3.34)	(3.57)	(3.57)	(3.54)	(3.54)	(3.37)	(4.66)	(4.12)
Byes	632	3.05	3.24	3.28	3.36	3.41	5.04	4.85	7.25
		(2.97)	(3.18)	(3.35)	(3.41)	(3.41)	(4.71)	(4.64)	(3.54)
Total	6,306	3.58	3.67	3.71	3.70	3.76	4.09	4.21	7.06
		(3.08)	(3.21)	(3.27)	(3.30)	(3.35)	(3.53)	(3.79)	(3.67)
Rank 6-15									
Wins	8,974	4.15	4.17	4.20	4.26	4.37	4.60	4.36	4.74
		(3.23)	(3.17)	(3.21)	(3.21)	(3.25)	(3.25)	(3.06)	(3.35)
Losses	2,436	2.43	2.48	2.56	2.59	2.62	2.80	4.31	3.64
		(3.47)	(3.57)	(3.62)	(3.67)	(3.68)	(3.73)	(2.76)	(2.65)
Byes	1,035	4.12	4.25	4.21	4.20	4.36	5.00	4.69	4.10
		(3.48)	(3.49)	(3.30)	(3.24)	(3.29)	(2.94)	(2.78)	(3.05)
Total	12,445	3.81	3.85	3.88	3.93	4.03	4.28	4.38	4.47
		(3.37)	(3.35)	(3.37)	(3.37)	(3.41)	(3.40)	(2.98)	(3.23)
Rank 16-25									
Wins	8,228	3.24	3.27	3.30	3.32	3.43	3.55	3.45	3.67
		(3.52)	(3.57)	(3.62)	(3.62)	(3.75)	(4.07)	(3.78)	(2.69)
Losses	3,179	1.19	1.19	1.20	1.19	1.19	1.21	1.63	3.12
		(2.42)	(2.44)	(2.44)	(2.43)	(2.45)	(2.49)	(1.81)	(1.86)
Byes	1,220	2.73	2.72	2.80	2.79	2.85	2.73	2.94	3.06
		(3.79)	(3.83)	(3.87)	(3.90)	(4.07)	(3.74)	(3.51)	(2.30)
Total	12,627	2.67	2.69	2.72	2.73	2.81	2.88	2.94	3.48
		(3.42)	(3.46)	(3.50)	(3.51)	(3.63)	(3.83)	(3.45)	(2.49)

Defn. 1, 2, 3, and 4 are $AggB$ ($AggW$) $= 1$ if aggregate rank is strictly greater than 1, 2, 3, and 4 ranks better (worse) than individual rank, respectively.

Non-social post is the estimated Bayesian posterior that only conditions on game scores (and not aggregates rank)

Table 2: Mean Rank Improvement (Prior Rank - Posterior Rank) by Prior Rank Group

	Definition 1				Definition 2			
	Wins		Losses		Wins		Losses	
Prior Rank Group	Observed	Estimated	Observed	Estimated	Observed	Estimated	Observed	Estimated
All Ranks								
Total	1.39	1.49	-4.89	-4.58	1.39	1.53	-4.89	-4.71
	(2.49)	(4.47)	(4.12)	(4.81)	(2.49)	(4.25)	(4.12)	(4.71)
	[21811]	[21811]	[6680]	[6680]	[21811]	[21811]	[6680]	[6680]
AggB	2.60	4.07	-3.35	-2.34	2.95	4.63	-2.67	-1.75
	(2.79)	(4.17)	(3.66)	(3.79)	(2.94)	(4.09)	(3.42)	(3.51)
	[4758]	[4758]	[1382]	[1382]	[2967]	[2967]	[838]	[838]
AggW	0.92	0.45	-4.73	-4.92	0.88	0.36	-4.36	-4.61
	(2.58)	(4.51)	(4.31)	(4.78)	(2.70)	(4.63)	(4.26)	(4.72)
	[6099]	[6099]	[2279]	[2279]	[4124]	[4124]	[1623]	[1623]
AggS	1.12	0.94	-5.71	-5.34	1.22	1.23	-5.53	-5.34
	(2.09)	(4.13)	(3.95)	(4.94)	(2.19)	(3.87)	(4.01)	(4.69)
	[10954]	[10954]	[3019]	[3019]	[14720]	[14720]	[4219]	[4219]
Ranks 1-5								
Total	0.06	-0.48	-7.65	-5.53	0.06	-0.29	-7.65	-5.88
	(1.29)	(2.03)	(3.60)	(5.71)	(1.29)	(1.85)	(3.60)	(5.74)
	[4609]	[4609]	[1065]	[1065]	[4609]	[4609]	[1065]	[1065]
AggB	0.55	0.81	-5.74	-0.94	0.78	0.91	-4.38	-0.15
	(1.50)	(1.57)	(2.45)	(3.16)	(1.74)	(1.50)	(2.57)	(3.21)
	[281]	[281]	[78]	[78]	[68]	[68]	[13]	[13]
AggW	-0.53	-2.29	-10.42	-7.86	-0.77	-2.89	-10.31	-10.38
	(1.84)	(2.27)	(5.11)	(4.89)	(2.22)	(2.90)	(5.23)	(5.16)
	[797]	[797]	[113]	[113]	[365]	[365]	[48]	[48]
AggS	0.16	-0.17	-7.46	-5.64	0.12	-0.08	-7.57	-5.74
	(1.06)	(1.74)	(3.25)	(5.76)	(1.14)	(1.53)	(3.45)	(5.67)
	[3531]	[3531]	[874]	[874]	[4176]	[4176]	[1004]	[1004]
Ranks 6-15								
Total	1.08	0.09	-7.29	-7.76	1.08	0.20	-7.29	-7.81
	(2.30)	(4.12)	(3.77)	(4.43)	(2.30)	(3.86)	(3.77)	(4.17)
	[8974]	[8974]	[2436]	[2436]	[8974]	[8974]	[2436]	[2436]
AggB	1.90	2.45	-5.87	-4.31	2.05	2.94	-5.49	-2.95
	(2.38)	(3.99)	(4.02)	(4.75)	(2.49)	(3.96)	(4.02)	(5.11)
	[1877]	[1877]	[488]	[488]	[1076]	[1076]	[244]	[244]
AggW	0.48	-1.42	-8.28	-9.51	0.38	-1.95	-8.52	-9.55
	(2.55)	(4.18)	(3.27)	(3.71)	(2.68)	(4.28)	(3.28)	(3.62)
	[2468]	[2468]	[813]	[813]	[1651]	[1651]	[527]	[527]
AggS	1.06	-0.05	-7.19	-7.99	1.10	0.29	-7.16	-7.97
	(2.01)	(3.69)	(3.79)	(3.89)	(2.08)	(3.36)	(3.75)	(3.62)
	[4629]	[4629]	[1135]	[1135]	[6247]	[6247]	[1665]	[1665]
Ranks 16-25								
Total	2.47	4.11	-2.12	-1.82	2.47	4.00	-2.12	-1.94
	(2.72)	(4.57)	(2.40)	(2.64)	(2.72)	(4.46)	(2.40)	(2.62)
	[8228]	[8228]	[3179]	[3179]	[8228]	[8228]	[3179]	[3179]
AggB	3.33	5.60	-1.61	-1.31	3.57	5.76	-1.45	-1.28
	(2.94)	(3.83)	(2.24)	(2.53)	(3.04)	(3.80)	(2.26)	(2.40)
	[2600]	[2600]	[816]	[816]	[1823]	[1823]	[581]	[581]
AggW	1.72	2.85	-2.12	-1.92	1.56	2.73	-2.00	-1.87
	(2.53)	(4.03)	(2.36)	(2.42)	(2.57)	(3.81)	(2.35)	(2.34)
	[2834]	[2834]	[1353]	[1353]	[2108]	[2108]	[1048]	[1048]
AggS	2.44	4.00	-2.54	-2.09	2.45	3.87	-2.45	-2.24
	(2.46)	(5.28)	(2.51)	(2.94)	(2.48)	(4.77)	(2.43)	(2.82)
	[2794]	[2794]	[1010]	[1010]	[4297]	[4297]	[1550]	[1550]

Standard Deviations in Parentheses; No of Observations in square brackets

Table 3: How do voters respond to social information?

	Wins: Definition 1			Wins: Definition 2			Losses: Definition 1			Losses: Definition 2		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Home	1.71***	1.71***	1.74***	1.60***	1.60***	1.63***	-1.09*	-1.09	-1.29**	-1.17**	-1.16*	-1.34**
	(0.25)	(0.28)	(0.27)	(0.24)	(0.28)	(0.25)	(0.56)	(0.68)	(0.59)	(0.55)	(0.70)	(0.57)
	[1.03]	[1.03]	[0.97]	[1.07]	[1.08]	[0.99]	[1.07]	[1.08]	[1.07]	[0.99]	[1.00]	[1.02]
Score margin	-0.062***	-0.062***	-0.065***	-0.060***	-0.060***	-0.063***	0.021	0.023	0.016	0.019	0.021	0.010
	(0.008)	(0.009)	(0.008)	(0.008)	(0.009)	(0.008)	(0.030)	(0.026)	(0.025)	(0.032)	(0.025)	(0.024)
	[0.68]	[0.68]	[0.68]	[0.69]	[0.69]	[0.68]	[0.91]	[1.03]	[1.63]	[1.78]	[2.54]	[0.61]
AggB	-1.78***			-1.78***			0.83**			0.98**		
	(0.13)			(0.16)			(0.39)			(0.38)		
	[0.52]			[0.48]			[0.26]			[0.25]		
AggW	0.08			0.34**			-0.49			-0.39		
	(0.15)			(0.15)			(0.39)			(0.35)		
	[0.83]			[0.74]			[1.07]			[0.54]		
Opprank * AggB		-1.45***			-1.30***			0.88			0.89*	
		(0.36)			(0.38)			(0.55)			(0.51)	
		[0.34]			[0.29]			[0.29]			[0.26]	
Opprank * AggW		0.07			0.73***			-0.74			-0.68	
		(0.35)			(0.26)			(0.52)			(0.53)	
		[0.09]			[92.12]			[2.59]			[2.05]	
Oppunrank * AggB		-1.83***			-1.86***			0.78			1.06*	
		(0.16)			(0.16)			(0.65)			(0.56)	
		[0.57]			[0.52]			[0.23]			[0.24]	
Oppunrank * AggW		0.07			0.28*			-0.26			-0.14	
		(0.16)			(0.17)			(0.54)			(0.46)	
		[0.31]			[0.53]			[0.42]			[0.12]	
Rank 1-5 * AggB			-0.73**			-0.48			3.47***			3.13
			(0.30)			(0.39)			(1.34)			(2.23)
			[0.49]			[0.31]			[0.68]			[0.51]
Rank 1-5 * AggW			1.23***			1.78***			0.31			-2.36
			(0.28)			(0.30)			(1.54)			(1.45)
			[0.78]			[0.76]			[0.13]			[0.48]
Rank 6-15 * AggB			-1.64***			-1.69***			2.51***			3.41***
			(0.25)			(0.23)			(0.60)			(0.92)
			[0.65]			[0.63]			[0.67]			[0.69]
Rank 6-15 * AggW			0.69***			1.44***			-0.13			0.06
			(0.21)			(0.22)			(0.44)			(0.42)
			[0.57]			[0.67]			[0.13]			[0.05]
Rank 16-25 * AggB			-0.76**			-0.81***			-0.01			0.20
			(0.32)			(0.27)			(0.28)			(0.29)
			[0.46]			[0.41]			[0.02]			[0.18]
Rank 16-25 * AggW			0.25			0.10			-0.29			-0.15
			(0.34)			(0.29)			(0.21)			(0.22)
			[0.29]			[0.11]			[2.82]			[0.51]
R-sq	0.110	0.110	0.200	0.106	0.107	0.197	0.094	0.095	0.171	0.101	0.101	0.166

Bootstrap standard errors clustered by game in parentheses. ***,**,* = significant at 1%, 5%, 10%.
Reaction ratios (described in the text) in square brackets. N = 21,811 for wins, 6,680 for losses.
Dependent Variable = $Over_{i,t}$. All regressions include voter, season and week FE, a constant, dummy for opponent's rank, interaction of this dummy with score margin.
Specifications that include interactions with Agg Worse and Agg Better also include those interactions as dummies in the regressions.

Table 4: Voter Characteristics Estimation Results: Wins

	Definition 1						Definition 2					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
			(X = Nat)	(X=Follow)	(X=Repeat)				(X = Nat)	(X=Follow)	(X=Repeat)	
Home	1.71***	1.66***	1.71***	1.71***	1.75***	1.71***	1.60***	1.54***	1.60***	1.60***	1.65***	1.60***
	(0.284)	(0.335)	(0.245)	(0.228)	(0.276)	(0.281)	(0.27)	(0.31)	(0.25)	(0.22)	(0.30)	(0.242)
	[1.03]	[1.05]	[1.03]	[1.03]	[1.04]	[1.03]	[1.02]	[1.05]	[1.02]	[1.02]	[1.05]	[1.02]
Score Margin	-0.062***	-0.061***	-0.062***	-0.062***	-0.062***	-0.062***	-0.060***	-0.058***	-0.060***	-0.060***	-0.062***	-0.060***
	(0.0102)	(0.00857)	(0.00949)	(0.0123)	(0.00938)	(0.00907)	(0.010)	(0.009)	(0.009)	(0.012)	(0.011)	(0.00801)
	[0.68]	[0.68]	[0.68]	[0.68]	[0.67]	[0.68]	[0.67]	[0.67]	[0.67]	[0.67]	[0.67]	[0.67]
AggB	-1.81***	-1.95***					-1.75***	-1.68***				
	(0.17)	(0.23)					(0.19)	(0.29)				
	[0.52]	[0.59]					[0.47]	[0.49]				
AggW	-0.07	0.31					0.21	0.70***				
	(0.19)	(0.22)					(0.19)	(0.27)				
	[2.97]	[3.56]					[0.53]	[1.18]				
Experience*AggB	0.01	0.010				0.01	-0.01	-0.003				-0.01
	(0.02)	(0.01)				(0.02)	(0.02)	(0.01)				(0.03)
	[0.39]	[10.30]				[0.43]	[1.04]	[0.44]				[1.21]
Experience*AggW	0.06**	-0.01				0.05***	0.05**	-0.02*				0.04**
	(0.02)	(0.01)				(0.02)	(0.02)	(0.01)				(0.02)
	[1.27]	[2.34]				[1.40]	[1.39]	[1.74]				[1.65]
X * AggB			-1.99***	-1.59***	-1.79***				-2.19***	-1.63***	-1.73***	
			(0.30)	(0.37)	(0.20)				(0.35)	(0.40)	(0.23)	
			[0.59]	[0.51]	[0.53]				[0.57]	[0.45]	[0.48]	
X * AggW			0.39*	0.48	0.01				0.60**	0.75*	0.22	
			(0.24)	(0.40)	(0.22)				(0.29)	(0.40)	(0.25)	
			[1.15]	[10.50]	[0.07]				[0.74]	[2.76]	[1.21]	
(1-X) * AggB			-1.76***	-1.78***	-1.87***				-1.75***	-1.79***	-1.96***	
			(0.15)	(0.15)	(0.23)				(0.18)	(0.15)	(0.21)	
			[0.52]	[0.53]	[0.54]				[0.48]	[0.49]	[0.52]	
(1-X) * AggW			0.05	0.06	0.322				0.31**	0.32*	0.72***	
			(0.15)	(0.17)	(0.22)				(0.15)	(0.17)	(0.22)	
			[0.68]	[0.61]	[0.70]				[0.69]	[0.64]	[0.72]	
2006 * AggB						-1.76***						-1.79***
						(0.24)						(0.27)
						[0.53]						[0.48]
2006 * AggW						0.23						0.61**
						(0.24)						(0.29)
						[0.46]						[0.58]
2007 * AggB						-1.94***						-1.72***
						(0.27)						(0.32)
						[0.50]						[0.44]
2007 * AggW						-0.27						0.01
						(0.28)						(0.24)
						[0.57]						[0.16]
2008 * AggB						-1.74***						-1.74***
						(0.24)						(0.30)
						[0.55]						[0.50]
2008 * AggW						-0.20						0.02
						(0.29)						(0.32)
						[1.34]						[0.11]
N	21811	15666	21811	21811	13394	21811	21811	15666	21811	21811	13394	21811
R-sq	0.11	0.11	0.11	0.11	0.117	0.111	0.106	0.105	0.106	0.106	0.113	0.107

Bootstrap standard errors clustered by game in parentheses. ***,**,* = significant at 1%, 5%, 10%.
Reaction ratios (described in the text) in square brackets
Dependent Variable: Over$_{i,t}$ = $r^B_{i,t} - r^O_{i,t}$ (Bayesian rank - Observed rank)
All regressions include voter, season and week FE, a constant, dummy for opponent's rank, interaction of this dummy with score margin.
Terms interacted with Agg Worse and Agg Better are also included as stand-alone regressors.

Table 5: Voter Characteristics Estimation Results: Losses

	Definition 1						Definition 2					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
			(X = Nat)	(X=Follow)	(X=Repeat)				(X = Nat)	(X=Follow)	(X=Repeat)	
Home	-1.09*	-1.03*	-1.09*	-1.09	-1.06**	-1.09	-1.17*	-1.11*	-1.17*	-1.17*	-1.11	-1.176**
	(0.647)	(0.606)	(0.627)	(0.682)	(0.535)	(0.681)	(0.65)	(0.59)	(0.62)	(0.67)	(0.82)	(0.552)
	[1.07]	[1.09]	[1.07]	[1.07]	[1.04]	[1.08]	[1.09]	[1.09]	[1.08]	[1.08]	[1.67]	[1.07]
Score Margin	0.021	0.02	0.021	0.021	0.022	0.022	0.019	0.02	0.019	0.02	0.02	0.0198
	(0.0292)	(0.0319)	(0.0340)	(0.0360)	(0.0290)	(0.0265)	(0.029)	(0.030)	(0.033)	(0.035)	(0.039)	(0.0301)
	[0.91]	[1.25]	[0.88]	[0.89]	[0.82]	[1.03]	[0.87]	[0.77]	[0.85]	[0.88]	[0.52]	[1.03]
AggB	0.72	0.10					0.87*	0.65				
	(0.44)	(0.74)					(0.45)	(0.60)				
	[0.23]	[0.27]					[0.21]	[0.19]				
AggW	-0.57	-0.25					-0.45	-0.49				
	(0.42)	(0.57)					(0.35)	(0.48)				
	[1.36]	[0.39]					[0.55]	[1.14]				
Experience*AggB	0.04	-0.003				0.04	0.04	0.02				0.05
	(0.05)	(0.02)				(0.05)	(0.04)	(0.02)				(0.04)
	[4.42]	[0.28]				[2.45]	[0.43]	[1.18]				[0.62]
Experience*AggW	0.03	-0.008				0.04	0.02	0.004				0.02
	(0.03)	(0.01)				(0.03)	(0.03)	(0.01)				(0.03)
	[2.69]	[1.85]				[1.94]	[0.85]	[0.26]				[1.24]
X * AggB			1.20*	0.89*	0.73				0.15	1.93*	0.95**	
			(0.72)	(0.52)	(0.54)				(0.68)	(0.99)	(0.47)	
			[0.38]	[0.28]	[0.25]				[0.07]	[0.39]	[0.26]	
X * AggW			-0.04	-0.67	-0.49				-0.16	-0.74	-0.37	
			(0.45)	(0.62)	(0.56)				(0.37)	(0.50)	(0.46)	
			[0.05]	[0.50]	[1.17]				[0.24]	[0.59]	[0.39]	
(1-X) * AggB			0.80**	0.83**	1.14**				1.04***	0.95**	1.01**	
			(0.34)	(0.41)	(0.49)				(0.33)	(0.39)	(0.48)	
			[0.25]	[0.26]	[0.31]				[0.26]	[0.25]	[0.25]	
(1-X) * AggW			-0.52*	-0.48	-0.11				-0.41	-0.37	-0.4	
			(0.30)	(0.34)	(0.42)				(0.28)	(0.29)	(0.38)	
			[1.26]	[1.21]	[0.12]				[0.54]	[0.51]	[0.67]	
2006 * AggB						1.18**						1.22**
						(0.60)						(0.61)
						[0.34]						[0.29]
2006 * AggW						-0.38						-0.42
						(0.48)						(0.59)
						[0.74]						[1.28]
2007 * AggB						0.73						0.93
						(0.76)						(0.74)
						[0.21]						[0.21]
2007 * AggW						-0.40						-0.42
						(0.70)						(0.65)
						[0.53]						[0.33]
2008 * AggB						0.09						0.11
						(0.67)						(0.58)
						[0.04]						[0.04]
2008 * AggW						-1.00						-0.54
						(0.63)						(0.59)
						[8.01]						[0.79]
N	6680	4752	6680	6680	4163	6680	6680	4752	6680	6680	4163	6680
R-sq	0.094	0.092	0.094	0.095	0.092	0.096	0.101	0.100	0.101	0.102	0.101	0.102

Bootstrap standard errors clustered by game in parentheses. ***,**,* = significant at 1%, 5%, 10%.
Reaction ratios (described in the text) in square brackets
Dependent Variable: Over$_{i,t}$ $= r^O_{i,t} - r^B_{i,t}$ (Observed rank - Bayesian rank)
All regressions include voter, season and week FE, a constant, dummy for opponent's rank, interaction of this dummy with score margin.
Terms interacted with Agg Worse and Agg Better are also included as stand-alone regressors.

Table 6: Robustness Checks

	Robustness Check 1				Robustness Check 2			
	Definition 1		Definition 2		Definition 1		Definition 2	
	Wins	Losses	Wins	Losses	Wins	Losses	Wins	Losses
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AggB	-1.74***	0.53	-1.36***	1.01*	-1.97***	1.09***	-2.05***	1.29***
	(0.28)	(0.76)	(0.39)	(0.59)	(0.13)	(0.40)	(0.16)	(0.40)
AggW	-0.45	-0.81	0.13	0.022	0.25*	-0.45	0.61***	-0.42
	(0.33)	(0.61)	(0.29)	(0.47)	(0.15)	(0.42)	(0.15)	(0.38)
N	7170	2010	7170	2010	21811	6680	21811	6680
R-sq	0.055	0.233	0.036	0.194	0.116	0.100	0.117	0.109

Bootstrap standard errors clustered by game in parentheses.
***,**,* = significant at 1%, 5%, 10%.
Dependent Variable: Over$_{i,t}$ = $r^O_{i,t} - r^B_{i,t}$ (Observed rank - Bayesian rank)
All regressions include all regressors included in model (1) of Table 3.
Robustness Check 1 = priors estimated with 2006 and 2007 seasons only; analysis conducted on 2008 season only.
Robustness Check 2 = priors estimated using aggregate final ranks as true ranks.

A Supplementary Appendix

Much of this material is drawn directly from Stone (2009).

A.1 Ranked and Unranked Teams

Most games are between ranked and unranked teams, since voters only rank 25 out of approximately 120 Division I-A teams. Consequently, some method of distinguishing among unranked teams is needed to account for heterogeneous quality among unranked teams. The method must rely on information observable to the voters, since we are assuming any information we condition our estimated posteriors on will also be used by the voters' to determine their posteriors. We use three primary variables for this purpose: 1) aggregate rank, if 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. For the first, we categorize aggregate rank in three bins; ≤ 25, 26-30 and 31 $\leq$. We also condition on year-to-date number of losses (0 versus >0 in weeks 1-3; 0-1 versus >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 83: three aggregate rank groups for teams ranked (by the particular voter) 2-25 or by at least one other voter and two aggregate rank groups for teams ranked number 1 by the voter (only two since these teams cannot have a "better" aggregate rank), so 77 rank groups for teams ranked by at least one voter, and six groups for unranked teams, those ranked in last two years, last five years, and not in last five years, with and without loss(es). This method of distinguishing among unranked teams is not sufficient for accurately estimating posterior beliefs for unranked teams. Consequently, we only estimate posterior beliefs and rankings for teams that are currently ranked. This forces a need to account for the fact that several teams do indeed drop from the rankings for most voters from each week to the next. We do this by restricting the maximum (worst) estimated posterior rank to be one greater than the number of teams that are observed to stay in the poll, by voter-week. We also re-rank observed posteriors among teams that were in the prior poll, and assign the same maximum rank (one greater than the number of teams that stay in the voter-week's top 25) to teams that drop from the observed rankings. This allows comparisons between estimated Bayesian and observed posteriors to be apples-to-apples, and unconfounded by teams entering the polls at various rank levels. In other words, it allows the estimated posteriors to potentially exactly match the observed posteriors. To illustrate this further by example, suppose only 22 of 25 teams in voter X's week 1 ballot are ranked by X in week 2. Suppose the teams ranked 19-21 in week 1 dropped out of X's top 25 and were replaced by new teams (teams unranked by X in week 1), so the ranks of teams ranked 1-18 and 22-25 did not change. Since we know relatively little about X's beliefs about the new teams in the poll (since they were unranked before they entered the poll) we ignore them and adjust the observed week 2 posteriors for X. We assign ranks 19-22 to teams with observed posterior ranks 22-25, and assign rank 23 to the teams that dropped out of X's ballot. For the Bayesian estimates, we assign rank 23 to all teams with estimated rank 23 or higher. Hence, the estimated rankings can potentially be exactly the same as the observed posteriors. Finally, due to the assumption that teams are ranked in order of expected rank, we have to calculate expected rank (for the estimated posteriors). Thus we have to use some value for expected rank conditional on being unranked. We use the value 35 for this purpose; results are similar with different values.

A.2 Score Distributions

The ideal way to estimate the score distributions (the g's) would be to use the historical distributions of scores between teams of the various final ranks by individual voters. We do not have historical individual final rank data, so we need to use the aggregate final rank data for this purpose. The other issue complicating the estimation of the g's is that although we have access to all historical scores, the sample sizes for scores between teams of particular ranks is highly limited. We use score data dating back to 1989 because that is when the AP Top 25 in its current form began.[30] 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, we 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 we are forced to use multiple smoothing techniques. First, we divide the 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 scores from games between teams in each category. Then, we divide the score distribution into categories (buckets) of size 7 (with upper and lower bounds of plus/minus 49+). We construct Bayesian posteriors using both the raw frequencies of scores in each bucket for games between teams in each rank group, and smoothed frequencies obtained using essentially a second-order moving average, and the differences are minimal. The smoothed estimates are referred to by default in the body of the paper, and the smoothing method is defined formally 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).[31] There are 17 points of support for each score margin distribution (-50-, [-49,-43],...,[-7,-1],0,[1,7],...,[43,49],50+). Let $c_{j,k}^i$ denote the historical count of games with score margins in category $i \in \{1, ..., 17\}$ 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 $\tilde{c}^i = \frac{1}{5}\Sigma_{\hat{i}=i-2}^{i+2} c^{\hat{i}}$. For $i \in \{1, 2\}$ let $\tilde{c}^i = \frac{1}{3+I(i=2)}\Sigma_{\hat{i}=1}^{i+2} c^{\hat{i}}$, in which $I(i = 2) = 1$ if $i = 2$, else $I(i = 2) = 0$. For $i \in \{16, 17\}$ let $\tilde{c}^i = \frac{1}{3+I(i=16)}\Sigma_{\hat{i}=i-2}^{18} c^{\hat{i}}$. $\tilde{c}^i = c^i$ if $i = 0$. Let $g(s_{jk}^i)$ denote the probability the score margin, s, is in category i for games between a home team in rank group j and away team in group k. Then the estimated probability, $\hat{g}(s_{jk}^i)$, is equal to $\frac{\tilde{c}^i}{\Sigma_{\hat{i}=1}^{17} \tilde{c}^{\hat{i}}}$.

[30] We use data from all regular season games but exclude games played at neutral sites. We 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). Since attempting to clean the data to account for injuries could create more noise than it would eliminate, we simply ignore the issue.

[31] We do not use the rank group 'unranked' but ranked by at least one other voter because there is no direct analog when using aggregate rankings. For example, in the aggregate rankings we observe the 26th ranked team (with the 26th most aggregate points). It is not clear an individual voter would consider a team he/she did not rank with the most votes by other voters as 26th best, however.

A.3 Prior Distributions

The prior distributions are estimated using the empirical final rank frequencies. A separate prior is estimated for each of the 83 prior rank groups or categories (see section A.1), for each week, but not for each season. For example, the first rank category is: teams ranked number one with the "same" aggregate rank (an aggregate rank not below three or four). To estimate the week 1 prior for this category, we look at the empirical frequencies with which teams in this category in week 1 finished in the various final rank groups (1-6, 7-12, 13-18, 19-25 and unranked; see section A.2). That is, to estimate the prior probability the team in this category has final rank 1-6, we take the number of teams in the category, for that week, that finish ranked 1-6 (by the same voter, in the same season), and divide by the total number of teams in the category, for that week. We do not need to do additional smoothing since sample sizes are much larger than for the score distribution estimation.

We estimate different priors for each week because it is natural to think priors become stronger as the season progresses and information is obtained. We could just estimate priors for week 1, and then use the posteriors we calculate based on those priors as the priors in subsequent weeks. We prefer to directly estimate priors for later weeks, because using our estimated posteriors as priors would introduce substantial noise. Since we have information on the prior-final rank relations for later weeks, it makes sense to use this information. Of course, we could condition the priors on other factors besides those described in section A.1, such as rank in the previous week. We ignore these because we believe any benefits they may offer would be outweighed by added computational and other costs.

It is important to note the priors are constant across voters in the sense that we assume that voters have the same priors, for a given (prior) rank and week. In other words, we assume each voter has the same belief that a team has true rank 1-6 for teams ranked number one with the same aggregate rank in week 1. This of course still allows for heterogeneity in prior beliefs, since the voters do in fact rank different teams in the different rank categories.