The Journal of Prediction Markets 2015 Vol 9 No 2 pp 1-31

A COMPARISON OF FORECASTING METHODS: FUNDAMENTALS, POLLING, PREDICTION MARKETS, AND EXPERTS*

Deepak Pathak[#] *David Rothschild* *Miroslav Dudík*
UC Berkeley *Microsoft Research* *Microsoft Research*
Berkeley, CA, USA *New York, NY, USA* *New York, NY, USA*
pathak@berkeley.edu davidmr@microsoft.com mdudik@microsoft.com

ABSTRACT

We compare Oscar forecasts derived from four data types (fundamentals, polling, prediction markets, and domain experts) across three attributes (accuracy, timeliness and cost effectiveness). Fundamentals-based forecasts are relatively expensive to construct, an attribute the academic literature frequently ignores, and update slowly over time, constraining their accuracy. However, fundamentals provide valuable insights into the relationship between key indicators for nominated movies and their chances of victory. For instance, we find that the performance in other awards shows is highly predictive of the Oscar victory whereas box office results are not. Polling-based forecasts have the potential to be both accurate and timely. Timeliness requires incentives for frequent responses by high-information users. Accuracy is achieved by a proper transformation of raw polls. Prediction market prices are accurate forecasts, but can be further improved by simple transformations of raw prices, yielding the most accurate forecasts in our study. Expert forecasts exhibit some characteristics of fundamental models, but are generally not comparatively accurate or timely. This study is unique in both comparing and aggregating four traditional data sources, and considering critical attributes beyond accuracy. We believe that the results of this study generalize to many other domains.

Keywords: prediction, Oscars, regression, probabilities, combining forecasts

---

* The authors would like to thank David Pennock for all of his help with this project, and Civic Science for providing the polling data.
# Corresponding author.

1 INTRODUCTION

The Oscars is the premier awards show of the American movie industry, watched live every year by millions around the world. Two of the more recent Oscars shows, in early 2013 and 2014, were watched live by over 40 million viewers and ads sold for between $1.8 million and $1.9 million per 30 seconds.[1] The 2013 and 2014 shows comprise the main set of outcomes discussed in this paper, each year consisting of awards in 24 categories ranging from some highly visible work, such as variations of the best picture and best actor awards, to some less visible behind-the-camera work, such as the best director and best cinematography awards. Similar to any highly popular live event, general public eagerly debates who will win the various categories. Movie studios wage campaigns for their movies, because Oscar victories provide new interest and revenue.[2] Thus, there is a strong interest and monetary incentive to forecast the winners accurately and see how the forecasts change over time. This paper examines several prominent data sources used to forecast the Oscar winners.

The goal of any forecast, what constitutes an efficient forecast, is to be *accurate*, *timely*, and *cost effective*. The forecast is *accurate* if it has a small error, but also if it is well calibrated and has an out-of-sample validity (i.e., it predicts the future rather than describing the past). The forecast is *timely* if it debuts early and updates often, so it is both fresh for stakeholders and granular for researchers to judge the impact of new information that is released during the entire evaluation period. For the Oscars, we start our forecasts at the release of the nominations, which is about six weeks before the show, and evaluate them daily. Finally, the forecast is *cost effective* if the insights gained justify the investment required to produce the forecast. Extending beyond the Oscars, a cost effective forecasting method should scale to many questions and domains.

The four data types discussed in this paper include fundamental data, polls, prediction markets, and domain experts. Fundamental data is "fundamental" because it is not created to answer our questions, but exists due to the nature of the event and the nominees. Examples of this data include the demographics of past winners or the box office receipts of nominated movies. We use the term *fundamentals* rather than the more common *quantitative* or *statistical models*, because we want to describe data, rather than the method of translating the data into a forecast. Unlike fundamentals, our other data sources (polls, prediction markets, and experts) are created to provide answers to the specific questions of interest. In polling, researchers ask respondents what they think or what they intend to do. Examples include polls asking

[1] http://variety.com/2013/tv/news/oscar-ad-prices-hit-all-time-high-as-abc-sells-out-2014-telecast-exclusive-1200778642/

[2] The average nominated movie for Best Picture now spends $10-15 million in their campaign: http://boxofficequant.com/the-value-of-an-oscar/

respondents which nominee they think will win certain categories. In prediction markets, traders can wager real or virtual money on the outcome of an event. For example, traders can buy and sell contingent contracts worth $1 if a nominee wins the Oscar and worth $0 if the nominee does not win the Oscar. The market prices of these contracts serve as the data for forecasting. Finally, domain experts publicly state their opinion on the likelihood of different outcomes. Examples of this data include movie columnists who state the probability that any given nominee will win selected categories.

Fundamental data can yield accurate forecasts in some domains (Fair 2011, Hummel and Rothschild 2014, Goel *et al.* 2010), but the fundamental models are costly to construct and generally not timely. Prior work has explored fundamental data's place in the timeline of the events, showing that it is most useful when less idiosyncratic information is available, as fundamental models are not good at absorbing dispersed or idiosyncratic information. This is true, for example, early in the election cycle in politics, before polling and prediction market data become available. Later in the cycle, the fundamental models fail to reflect available information and do not update in a timely manner relative to other data types (Lock and Gelman 2010, Rothschild 2015). In this paper, fundamental data includes box office returns, the number of theater screens showing the movie, ratings, and results in other awards, all of which we use to construct statistical models. Creating fundamental models is generally expensive due to the effort to identify suitable data sources, which tend to be highly question-specific. The modeling is further complicated by the fact that some of the data is not available for the full time frame, which either requires creating a separate model for each time frame, or a more sophisticated modeling approach.

Polling data can create accurate, but not necessarily timely forecasts. There is a vast literature showing that random and representative polling creates accurate forecasts of upcoming events (Erikson and Wlezien 2008). This literature is generally skeptical of non-random or non-representative polling (Squire 1988). However, the modest but growing literature (e.g., Ghitza and Gelman 2013, Wang et al. 2015) is beginning to show the value of non-representative polls. For instance, Rothschild and Wolfers (2011) demonstrate empirically how non-representative polling can benefit from asking more appropriate questions, such as the expectation questions for aggregate forecast. This paper contributes to the non-representative polling literature. The polling data we test is the expectation poll, in a selection of categories, administered to both self-selected and random (but non-representative) respondents. Similar to standard polling, our data is going to be most accurate after it is collected. Timeliness suffers since polls rarely collect consistent responses day-by-day. Cost is also a concern since the recruitment of respondents for standard representative polls costs tens to hundreds of dollars per respondent, and even non-representative polls require some advertising space or other active recruitment effort.

Prediction markets are accurate, timely, and cost effective. There is a growing literature on the efficiency of prediction markets in general (Arrow *et al.* 2008, Wolfers and Zitzewitz 2004) and in the movie industry in particular (Pennock *et al.* 2001). We examine three different sets of prediction market data: two real-money markets and one play-money market. The decision to include the play-money market was motivated by prior work showing their efficiency, including the market studied in this paper (Pennock *et al.* 2001).

The accuracy of expert forecasts can suffer due to the incentives of the forecasters, and timeliness is typically not a key priority. For instance, experts may suffer from herding and over-reliance on within-sample evaluation (Guedj and Bouchaud 2005). Many experts are not incentivized to provide the most accurate forecasts: an expert may place his or her forecasts near the center of other forecasts to avoid making a distinct mistake, or may place their forecast at the edge of other forecasts to achieve big wins (Hong *et al.* 2000). Further, with a few exceptions, such as earning-per-share estimates, experts tend to provide just one forecast before an event, rather than continuously update their forecasts as new information arrives.

This paper presents three contributions. First, we compare and aggregate four traditional data sources that are rarely analyzed together. We show how to utilize these diverse data sources individually, compare the resulting models, and consider benefits of aggregation. We ultimately provide a simple and reusable translation methodology for creating forecasts from raw fundamental, polling, and prediction market data. Beyond the literature on the four individual data types, our analysis also adds to the literature on forecast methodology with multiple data types (Clemen and Winkler 1986, Diebold and Mariano 2002, Harvey *et al.* 1998). Our second contribution is the focus on multiple critical attributes of forecasts, beyond accuracy, that are key for practitioners and underappreciated in the academic literature. By focusing on multiple attributes, we obtain a more nuanced and comprehensive evaluation. While we gain some interesting domain-specific insights from the fundamental data, we find that the prediction market data yields the most accurate and timely forecasts at scalable costs, whereas polling can also perform well under the right conditions. Third, we provide an in-depth study of forecasting in the movie industry, a domain which is underexplored in the academic literature despite its business relevance and high degree of public interest. Our study is possible thanks to a unique dataset of consistent polling outside of politics. While the paper focuses on a specific domain, we believe that the insights and methods generalize to many other domains.

2 DATA, ESTIMATION STRATEGY, AND RESULTS

2.1 TARGET DOMAIN: THE OSCARS

The main outcome variables for this paper are the 24 categories of the Oscars awarded on February 24, 2013 and March 2, 2014. All but one

category, Best Picture, comprise five nominees. The number of nominees in the Best Picture category can vary from year to year, but in each of the years 2013 and 2014 the category had nine nominees. The winner of each category is determined by the largest number of votes among the approximately 6,000 members of the Academy of Motion Picture Arts and Sciences. The same set of voters also decides on the initial nominations, although for some categories the nominations are decided by a subset of the voters. Their names and demographics are only partially known.[3]

2.2 NOTATION AND METRICS

Categories are indexed as i, nominees within each category as j, the final outcome is denoted Y_{ij} and is equal to 1 if the nominee j wins the category i, and zero otherwise. Forecasts are real-valued numbers p_{ij} predicting the probability of the j^{th} nominee winning the category i. We measure the accuracy of forecasts for category i by root mean squared error (RMSE):[4]

$$\mathrm{RMSE}(i) = \sqrt{\frac{1}{m_i}\sum_{j=1}^{m_i}\left(p_{ij} - Y_{ij}\right)^2}$$

where m_i is the number of nominees in category i. The performance of a forecast across several categories, say categories in a set I, is measured by an average RMSE:

$$\mathrm{RMSE}(I) = \frac{1}{|I|}\sum_{i \in I} \mathrm{RMSE}\,(i)\,.$$

2.3 FUNDAMENTAL DATA

Fundamental data is data that researchers do not necessarily collect to answer a forecasting question; the data exists for other reasons. When forecasting with fundamental data, this data is used to fit a statistical model to answer a question of interest. Collection of fundamental data is costly in domains like the Oscars where each of the 24 categories requires its own domain-specific data. Further, not all data is available at all time points during the forecasting period. For example, the outcomes of the preceding awards

[3] The L.A. Times was able to contact what they believed was 88% of the Academy voters in 2012: http://www.latimes.com/entertainment/la-et-movie-academy-methodology-html-htmlstory.html

[4] RMSE is a variant of Brier score, which is a standard accuracy measure for probabilistic forecasts.

shows in the same season (e.g., Golden Globes) prove to be highly predictive, so it is critical to include them in the model as soon as they become available. With the data in hand, we can construct statistical models. In our case, we forecast the probabilistic outcome using jointly determined models for all instants in time across the 24 different categories; with 24 categories and six distinct periods of data availability, we have 144 different models. In this section we describe our fundamental data, derive the fundamental models, and present cross-time comparisons of accuracy.

Most of the natural candidates for the fundamental data for movies, such as box office receipts, ratings, etc., reflect a holistic view of the movie, so they provide little predictive power for categories that focus on specific attributes. To overcome this limitation, we supplement the holistic data, wherever possible, with category-specific data. The most prominent of this is the data from prior awards shows in the same season. Our initial pool of fundamental data includes the following data sources for all nominations in years 1978—2014: name of person (if applicable), gross revenue and theater screens per week for the first eight weeks of movie release, release date, critical and popular rating, MPAA rating, genre, and budget. Further we include nominations and winnings in the prior awards shows as they unfold in the season: Critics' Choice Award, Golden Globes, Guild Awards, British Oscars, and Spirits Awards. The box office data is taken from Box Office Mojo (boxofficemojo.com), the records for awards shows from IMDb (imdb.com), and rankings from Rotten Tomatoes (rottentomatoes.com).

The fundamental data described in the previous paragraph can be incorporated in models in many different ways. To limit the set of possibilities, we use simple tests of predictive power to scale down the number of variables and determine the most predictive variable transformations. For instance, despite collecting data from 1978 onward, we only use the data from 1995 on, because the earlier data makes the forecasts less accurate. We translate the release date as the number of days before the night of the Oscars (e.g., February 25, in 2013). We include critical and popular ratings from Rotten Tomatoes in their original form as numeric variables scaled [0,100]. The box office data consists of weekly gross revenue and number of screens for the first eight weeks after release. This data is transformed into four variables. First two variables are obtained by fitting a linear model to the revenue per screen over time—the resulting constant and slope of the linear model constitute the first two derived variables. The second pair of variables is based on the movie's "wide opening week", defined as the week when the movie is shown on more than 600 screens in the United States and Canada. The two variables are the revenue per screen and the index of the week (counted backward starting from the Oscar night). The monetary values are scaled to current currency. We represent the performance in each of the preceding awards shows, namely, Critics' Choice Award, Golden Globes, Guild Awards, British Oscars, and Spirits Awards, by four variables: the indicator of a nomination and a win in the corresponding category, as well as

the total number of nominations and wins in a given awards show. This is a non-trivial exercise as categories are not well aligned among the awards shows and change over time, so the matching needs to be done by hand. For example, in the Golden Globes, the Best Motion Picture awards are given separately for Drama, and Musical or Comedy, and we consider both equally relevant.

The variables are indexed by $k = 1 \ldots d$. The year of the competition is denoted t. Outcomes in year t are denoted Y_{tij}. The value of the k^{th} variable for the category i and nominee j in year t, is denoted X_{tijk}. Apart from the original variables (described in the previous paragraph) we introduce additional variables to represent missing data. Specifically, for each original variable X_k which has some missing entries, we fill those missing entries with 0, and introduce a new variable, equal to 1 whenever X_k is missing, and zero otherwise. This modeling approach corresponds to the assumption that the stochastic pattern of missingness is the same during data collection as during the model evaluation.

We construct different models for each of the six evaluation periods. When fitting the model, we drop years 2013 and 2014. The data from 2013 and 2014 is only used for out-of-sample evaluation of accuracy.

Our modeling proceeds in two steps. In the first step, we fit a logistic model separately for each nominee and each category. We assume that the parameter vector is shared across nominees, i.e., we model the probability $\tilde{p}_{tij}$ of the nominee j winning the category i in year t as

$$\log \frac{\tilde{p}_{tij}}{1 - \tilde{p}_{tij}} = \boldsymbol{\beta}_i \,.\boldsymbol{X}_{tij}$$

where $\boldsymbol{X}_{tij}$ is the vector of variables X_{tijk} (across all k). In the second step, forecasts $\tilde{p}_{tij}$ are rescaled to sum to one within each category, yielding the final forecasts p_{tij}.

To obtain $\tilde{p}_{tij}$, we fit the models separately for each category by L1-penalized log likelihood, also known as *lasso* (Tibshirani 1996):

$$\widehat{\boldsymbol{\beta}}_i = \underset{\boldsymbol{\beta}_i}{\operatorname{argmax}} \left\{ \sum_{t,j} \left[Y_{tij} \log \tilde{p}_{tij} + \left(1 - Y_{tij}\right) \log\left(1 - \tilde{p}_{tij}\right) \right] - \lambda \sum_{k=1}^{d} |\beta_{ik}| \right\}$$

where the regularization coefficient λ is chosen by fivefold cross-validation. The motivation behind using the L1 penalty is that we expect that many

variables are irrelevant, so the solution $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ has many zero entries, which is encouraged by L1 penalty (Tibshirani 1996).

The above procedure yields 144 models, six for each of the 24 categories. Instead of listing coefficients of all models, we only provide the coefficients of the final time-slice model of all 24 categories (Tables 4–7 in Appendix A) and all six time-slice models for Best Picture (Table 1 below). We highlight a few key observations. First, as shown in Table 1 with large coefficients for the Golden Globes and BAFTA, the awards shows have most of the predictive power. This is especially true outside of the Best Picture and Best Director categories since other variables have only little predictive power there (see Tables 4–7). With that in mind, we make a few additional observations. Most of the time, the ratings (critical or popular) have no predictive power, but in a few cases (Original Screenplay in Table 5 and Documentary Short in Table 7) it is the popular ratings that predict the Oscar. The release date does matter, but in an unexpected manner. Note that we forecast an Oscar victory conditional on being nominated. While a movie is more likely to get a nomination if it opens later in the season, conditional on being nominated, the early-release movies are a little more likely to win. Conditional on being nominated, there is very little predictive power from the success in box office as shown both in Table 1 and Tables 4–7.

The six models for Best Picture demonstrate the evolution of the forecasts through the awards season. As Table 1 shows, more and more variables become available as we move in time from left to right, with the first model applicable at the moment when nominations are released and the last applicable just days before the Oscar night. The overall number of Oscar nominations is always predictive; this is not surprising as the Oscar voters are likely to think well of the decisions of other Oscar voters, who not only vote for the winners, but also choose the nominees. While awards show nominations are available from the beginning, the wins only become available after the date of the corresponding show. The coefficient of the wins can be an order of magnitude larger than the coefficient of the corresponding nomination. The first four awards shows have similarly large coefficients, whereas the fifth show (Spirit Awards), which is aimed towards independent films, provides little additional predictive power.

We evaluate the accuracy of our models by the average RMSE across all 24 categories. Figure 1 compares our models with the random forecast $p_{ij}^{\text{random}} = \frac{1}{m_i}$ where m_i is the number of nominations in the category. The dotted vertical lines represent the declaration of results for other awards shows in following order: Critics' Choice, Golden Globes, Guild, BAFTA, and Spirits. First, notice that we improve over the random forecast even before the results of the first awards show are announced. This means that the nominations in other awards shows, without the awards yet, and non-award

Table 1. Coefficients for all six models for Best Picture. Model 1 is the first model in time, determined from the data available at the nomination. Each successive model is available at a later point in time, until Model 6, which is the last model that can be determined just a few days before the Oscar night. Standard errors are provided in parentheses. Dashes indicate that the coefficient and its standard deviation are zero. Asterisk indicates that the variable was not available when the corresponding model was constructed.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Constant						
Gross/Screen	------	------	------	------	------	------
Slope Gross/Screen	------	------	------	------	------	------
Week Wide	0 (0.056)	0 (0.046)	0 (0.053)	0 (0.056)	0 (0.051)	0 (0.051)
Gross/Screens Wide	------	------	------	------	------	------
Release Date	0.002 (0.002)	0 (0.002)	0.001 (0.002)	0.001 (0.002)	0.002 (0.002)	0.002 (0.002)
Popular Rating	0 (0.016)	0 (0.010)	0 (0.013)	0 (0.013)	0 (0.014)	0 (0.014)
Critical Rating	0 (0.011)	0 (0.011)	0 (0.012)	0 (0.013)	-0.006 (0.017)	-0.006 (0.017)
Oscar Overall Nom	0.258 (0.131)	0.273 (0.121)	0.258 (0.119)	0.271 (0.121)	0.243 (0.109)	0.243 (0.109)
Critics Overall Nom	0 (0.030)	0 (0.024)	0 (0.030)	0 (0.034)	0 (0.020)	0 (0.020)
Critics Overall Win	*------	0.057 (0.126)	0.012 (0.112)	0 (0.103)	0 (0.054)	0 (0.054)
Critics Category Nom	0 (0.212)	0 (0.330)	0 (0.371)	0 (0.391)	0 (0.383)	0 (0.383)
Critics Category Win	*------	1.653 (0.750)	1.480 (0.791)	1.517 (0.796)	1.127 (0.823)	1.127 (0.826)
GG Overall Nom	0.052 (0.097)	0.001 (0.074)	0 (0.053)	0 (0.055)	0 (0.060)	0 (0.061)
GG Overall Win	*------	*------	0.282 (0.225)	0.278 (0.220)	0.232 (0.202)	0.232 (0.202)
GG Category Nom	0 (0.144)	0 (0.094)	0 (0.143)	0 (0.168)	0 (0.201)	0 (0.199)
GG Category Win	*------	*------	0 (0.172)	0 (0.153)	0 (0.148)	0 (0.148)
Guild Overall Nom	0.363 (0.208)	0.312 (0.213)	0.294 (0.212)	0.202 (0.194)	0.209 (0.190)	0.209 (0.190)
Guild Overall Win	*------	*------	*------	0.367 (0.302)	0.376 (0.301)	0.376 (0.301)
Guild Category Nom	------	------	------	------	------	------
Guild Category Win	*------	*------	*------	------	------	------
BAFTA Overall Nom	0.092 (0.076)	0.045 (0.065)	0.042 (0.065)	0.039 (0.062)	0 (0.019)	0 (0.019)
BAFTA Overall Win	*------	*------	*------	*------	0.357 (0.189)	0.357 (0.189)
BAFTA Category Nom	0.002 (0.353)	0 (0.200)	0 (0.222)	0 (0.263)	0 (0.256)	0 (0.255)
BAFTA Category Win	*------	*------	*------	*------	0 (0.364)	0 (0.363)
Spirit Overall Nom	0 (0.081)	-0.051 (0.092)	-0.064 (0.102)	-0.046 (0.094)	-0.040 (0.086)	-0.040 (0.078)
Spirit Overall Win	*------	*------	*------	*------	*------	0 (0.045)
Spirit Category Nom	------	------	------	------	------	------
Spirit Category Win	*------	*------	*------	*------	*------	------
Constant	-5.609 (1.790)	-5.207 (1.680)	-5.443 (1.899)	-5.641 (1.986)	-5.265 (2.264)	-5.265 (2.263)

variables do provide some information. Second, the biggest drops in error occur after the Golden Globes and the BAFTA awards, suggesting that they provide the most accurate signals. This is relatively consistent across many categories, as shown in Tables 4–7 in Appendix A. Finally, we show 2014's results in Appendix B as Figure 10. The awards shows do not happen on the same day, but there are still the same sharp increases in accuracy when the major awards are announced.

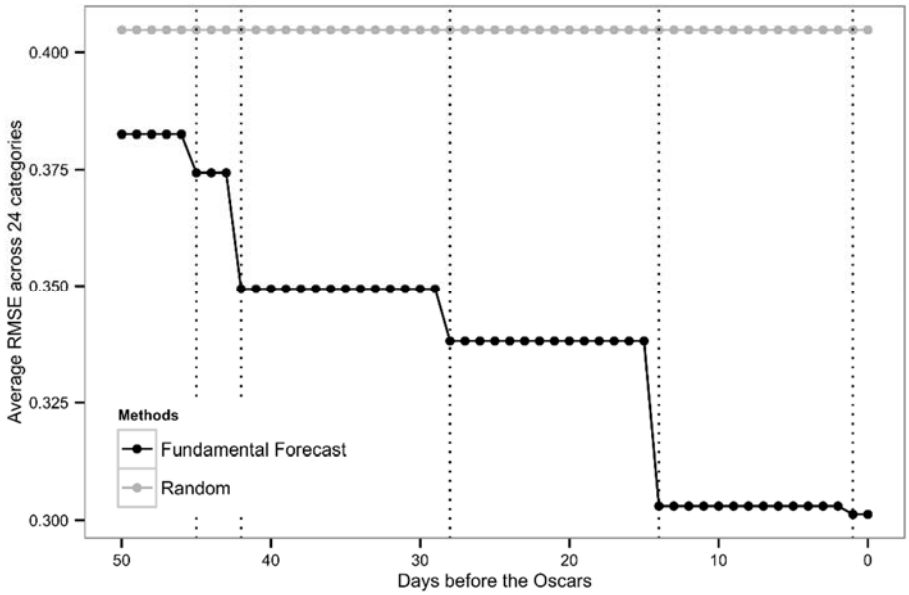

Figure 1. ***Average RMSE of fundamental model across all 24 categories*****. Out-of-sample error across all 24 categories for the 2013 Oscars. Prior awards shows are represented by vertical dotted lines, from left to right: Critics' Choice, Golden Globes, Guild, BAFTA, and Spirits.**

2.4 POLLING

Since the inception of representative polling in the 1930's, polling has been a central data type in forecasting upcoming events. We explore two different types of *non-representative* polling. Our polling data comes from a study by Civic Science,[5] which conducted online public polling across nine categories for 40 days before the Oscars in 2013. Motivated by the previous research indicating that expectation questions work well in non-representative polls (Rothschild and Wolfers 2011), Civic Science asked online users who they expected to win the Oscars. Two separate user populations were surveyed. First, the "random population" (but non-representative) was chosen randomly among the organic visitors of websites across the country that use polling powered by Civic Science. These respondents were asked for expectations in up to nine specified categories. Second, the "self-selected population" provided expectations in up to seven specified categories directly on Civic Science website. The self-selected respondents came from two groups. The first group came as a result of a massive social media push about 20 days prior to the Oscars. The second group consisted of regular visitors of

[5] Civic Science is contracted by over 500 newspaper websites and blogs across the country to conduct polling and subsequent data intelligence.

the Civic Science website. Note that both the "random" and the "self-selected" respondents are highly non-representative and there was no effort to make them representative of the population of the Academy voters.

We first consider a naive strategy that treats the polled fractions of the individual nominees as forecasts of their winning. Then we show that the accuracy of these naive forecasts can be improved by using a suitable transformation, which we call "translation". Finally, we compare the accuracy of our two polled populations.

Our comparison of the two populations has some limitations. First, they answer in bulk at different times. Figure 2 shows that the self-selected answers bunch early in the evaluation period, mainly around the social media push to gather respondents, while random respondents bunch mainly late in the evaluation period, when the Oscar questions were more frequently shown on the partner websites. There is no time period where both populations show high activity. Second, the self-selected respondents were only provided seven questions, while the random respondents saw nine.

The number of responses each day is neither consistent nor sufficient for accurate results, so we do not have a forecast every day for each type of poll. To maintain consistency in the analysis of results, each day is treated as a separate forecast using the polls from that day only. Days are indexed by t. Let $\tilde{p}_{ij}(t)$ denote the fraction of polls in i^{th} category supporting the j^{th} nominee; we call $\tilde{p}_{ij}(t)$ raw polls. Using raw polls as forecasts is a standard practice in politics and many other domains. In order to keep the magnitudes of RMSE comparable across categories, we standardize our data so that each category is treated to have five possible outcomes. For all categories except the Best Picture, the five outcomes correspond to the actual nominees. In the Best Picture category, which has nine nominees, we keep the top four polling nominees on a given day as separate outcomes and merge the bottom five into a single pseudo-nominee.[6]

The "Raw Fraction" lines in Figure 3 compare the average RMSE across seven common categories at any time t, for both self-selected and random respondents. We can see that the forecasting error of the random respondents' expectation decreases with time, as the number of respondents per day and the available information about the Oscar nominees increases. Self-selected respondents have a much smaller error on their first day, 20 days before the Oscars, than the random respondents ever achieve. Over time this lead fades as the few people who trickle to the Civic Science website are not as knowledgeable as those that answered the poll on the first day. Recall that the initial group was the result of a targeted social media push, whereas the remaining days consist of the regular traffic to the Civic Science website, comprising users who want to answer general polling questions. The fact that the initial group of self-selected respondents achieves a lower error 20 days

[6] The group of top four was consistent for all except two days of our evaluation period in 2013.

before the Oscars than a comparable number of random respondents a few days before the Oscars suggests that the self-selected respondents have more information per user. However, since they are more invested in the domain, they are also more likely to have strong preferences, which may lead to the "wishful thinking" bias (Granberg and Brent 1983). The random polling benefits from a larger number of respondents, demonstrating that a sufficient number of polls among random respondents can provide a meaningful result.

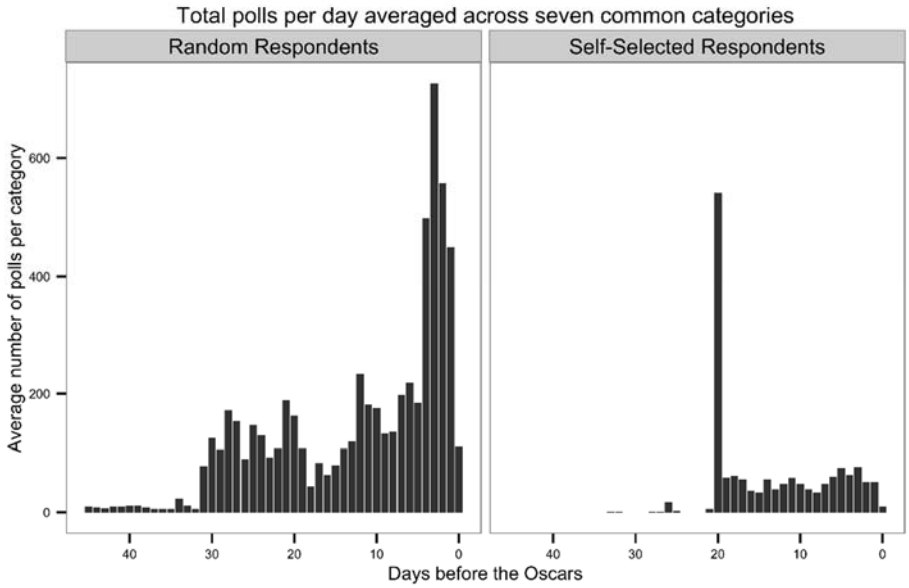

Figure 2. ***Votes per day for random and self-selected respondents in polling.*** **Averaged across the seven categories presented to both self-selected and random respondents.**

While the accuracy of raw polls is substantially better than that of a random forecast, it has been noted that it can be further improved by applying a suitable transformation (e.g., Erikson and Wlezien 2008). We now introduce and evaluate one type of a transformation, which we call "translation". It is motivated by two plausible conditions. First, the forecast should give a probability distribution among nominees (i.e., $\sum p_{ij} = 1$). Second, we aim to create a forecast that maintains the ranking of the nominees, i.e., a nominee that polls at 40% should be forecasted to win with a higher probability than a nominee that polls at 25%.

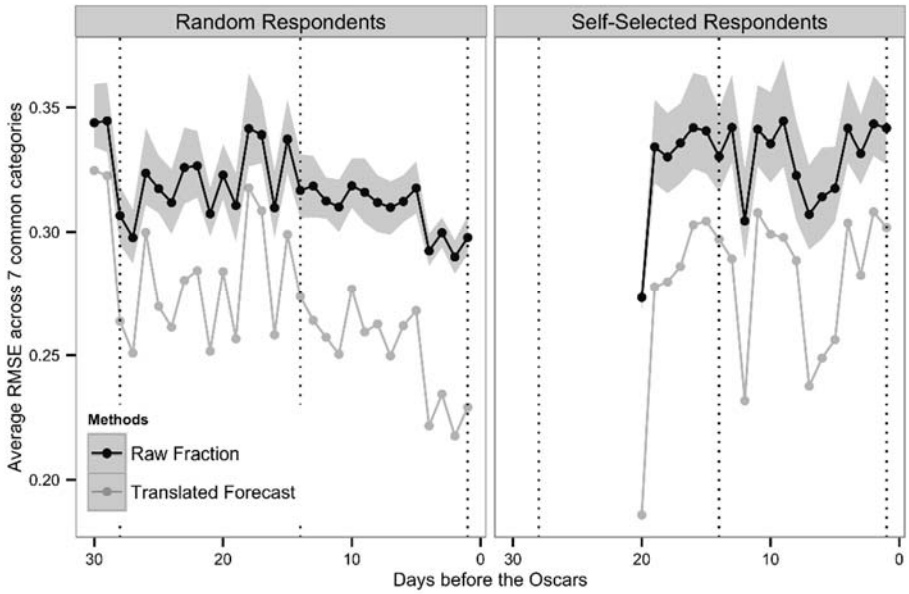

Figure 3. ***RMSE of raw and translated polls across seven common categories*****. Average error across the seven categories presented to both self-selected and random respondents for the 2013 Oscars. Prior awards shows are represented by vertical dotted lines, from left to right: Guild, BAFTA, and Spirits.**

Let $n_{ij}(t)$ be the number of respondents voting for the j^{th} nominee and $n_i(t)$ be the total number of respondents for category i on day t. Recall that m_i is the number of nominees in category i (in our case it is always 5). Our *translated forecast* is of the form:

$$p_{ij}(t) = c_i(t)\left(\frac{n_{ij}(t)+1}{n_i(t)+m_i}\right)^{\beta(t)} .$$

The leading term $c_i(t)$ is chosen to ensure that probabilities sum to one. The fraction $\frac{n_{ij}(t)+1}{n_i(t)+m_i}$ is a smoothed version of the raw poll $\tilde{p}_{ij}(t) = n_{ij}(t)/n_i(t)$ and it differs by including a "pseudo-vote" for each nominee (the approach known as Laplace smoothing). Finally, $\beta(t)$ is a time-varying function that parameterizes how much we want to exaggerate (if $\beta(t) > 1$) or diminish (if $0 < \beta(t) < 1$) the smoothed raw polls. The function $\beta(t)$ is the only unknown component of the model since the normalizer $c_i(t)$ is determined once $\beta(t)$ is. We parameterize it as a linear function of time:

$$\beta(t) = \beta_0 + \beta_1 t \ .$$

The solution is obtained by maximizing log likelihood with a ridge penalty on β_1, keeping β_0 unpenalized:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \sum_{i,t} \sum_{j} Y_{ij} \log p_{ij}(t) - \lambda \beta_1^2 \right\} .$$

The regularization coefficient λ is chosen by fivefold cross-validation. The best cross-validated log likelihood is achieved in the limit $\lambda = \infty$, i.e., the linear term does not yield additional explanatory power, and the best fit is obtained by the constant function $\beta(t) = \beta_0$.

The resulting values of β_0 are 2.03 ± 0.04 for self-selected respondents, 1.82 ± 0.05 for random respondents, and 1.92 ± 0.08 for all respondents.[7] Thus, in both considered datasets, the most accurate predictions are obtained by applying fairly large translation coefficients. Our raw polling data, which represents the ratios of respondents that expect a nominee to win, is too moderate when taken as a probability of that nominee winning or losing.

Figure 3 shows that the translation of raw polls, using our methodology, gives significantly more accurate forecasts; this result holds with out-of-sample 2014 data as well (not shown), though the difference is not as large.[8] Raw polls are plotted with [5-95]% confidence interval obtained by bootstrapping (over the poll responses). With a translation, the random respondents eventually create a similar error as the initial burst of self-selected respondents at 20 days before the Oscars, but not until the last few days of the evaluation period. The translated results for 2013 should be interpreted with caution since the reported accuracy is within the same sample that we use to estimate the translation parameters β_0 and β_1; even though the 2014 improvements are more moderate, overall these results show the promise of our translation methodology for polling.

2.5 PREDICTION MARKETS

In prediction markets, users can buy and sell contracts (securities) whose value is contingent on the outcome of an upcoming event. The price of the security is suggestive of the probability of the outcome. For example, the 2014 Oscar season included a security for Daniel Day Lewis to win Best Actor that would be worth $1 if he won and $0 if he lost. Since he was extremely likely to win, people were willing to pay nearly $1 for the security, demonstrating their subjective probability was approaching 100%. In this

[7] Standard errors obtained by cross-validation.

[8] We have verified that the decreased benefit of translation for 2014 is observed regardless whether the translation coefficients are fit using 2013 data (i.e., out-of-sample) or using 2014 data (i.e., in-sample) and the effects of translation are similar in both cases.

paper we consider three prediction markets, Betfair, Intrade and HSX, and the prediction market aggregator PredictWise. We first compare how they influence each other, then examine the forecasting accuracy of raw prices, and finally show the benefits of applying a translation correction similarly to polling.

We begin by describing the market aggregator PredictWise which published its forecasts live during the Oscar season. Let $\tilde{p}_{ij}^{\text{Betfair}}$ and $\tilde{p}_{ij}^{\text{Intrade}}$ denote the raw prices, in the two respective markets, of the security for the j^{th} nominee in i^{th} category (at a particular instant of time). The PredictWise forecast is derived as:

$$p_{ij}^{\text{PredictWise}} = c_i \Phi\left(\beta \Phi^{-1}\left(\frac{\tilde{p}_{ij}^{\text{Betfair}} + \tilde{p}_{ij}^{\text{Intrade}}}{2}\right)\right)$$

where Φ is the probit link, the scalar c_i ensures that the forecasts in the same category sum to one, and the parameter β plays the same role as $\beta(t)$ in the translation of polling. Specifically, the model begins with the average of raw prices from Betfair and Intrade, and then exaggerates extreme probabilities by applying $\beta > 1$; in our case, $\beta = 1.32$. This ex-ante model is inspired by a model of presidential prediction market (Rothschild 2009, 2015).

In 2013, Betfair traded 24 securities corresponding to all the categories, while Intrade and HSX had six and eight securities respectively; when Intrade data was not available, PredictWise used just the raw Betfair prices instead of the average of Intrade and Betfair. The data starts on the Monday after the announcement of the Oscar nominations and is recorded once per day at 11 PM ET. The prediction markets were relatively consistent with each other in terms of security price variation over time. Table 2 illustrates the relationship between the current prices on each exchange and the previous day's prices on other exchanges. In the first two rows we show that Betfair and Intrade are statistically predictive of each other's prices. The middle two rows show that, while technically statistically significant for Betfair, there is only very small additional predictive power in HSX's earlier prices that is not already captured by Betfair's or Intrade's prices (and similarly vice versa, as the last two lines show).

Interpreting raw security prices as forecasts has been widely studied (Manski 2006, Wolfers and Zitzewitz 2006). Figure 4 shows the average RMSE of raw prices across six common categories in Betfair, Intrade, HSX and PredictWise. First, note that errors of all markets decrease towards the Oscars night. Second, there are big drops in errors around the third and fourth awards show. Third, the two real-money prediction markets are in virtual lock-step, which indicates market efficiency. Fourth, the real-money markets have much smaller errors than the play-money market. In the rest of the paper, we use Betfair as our main prediction market data source, because it essentially agrees with Intrade over categories considered by Intrade, but

exists in all 24 categories, whereas it is much more accurate than HSX, both in raw prices and when translated into forecasts. Fifth, PredictWise, which can be viewed as a translated version of the Betfair+Intrade average, has a lower error, thus confirming our desire to translate raw prices.

Table 2. ***Relationship between Betfair, Intrade, and HSX prediction market data with a one-day lag*****. Statistically significant coefficients (at 1%) are denoted by *. Standard errors are provided in parentheses.**

Regression formula	Coefficient β	Coefficient γ
$\text{Betfair}_{t+1} = \alpha + \beta\text{Betfair}_t + \gamma\text{Intrade}_t + noise$	0.90 (0.03)*	0.10 (0.03)*
$\text{Intrade}_{t+1} = \alpha + \beta\text{Intrade}_t + \gamma\text{Betfair}_t + noise$	0.80 (0.03)*	0.23 (0.03)*
$\text{Betfair}_{t+1} = \alpha + \beta\text{Betfair}_t + \gamma\text{HSX}_t + noise$	0.99 (0.00)*	0.02 (0.01)*
$\text{Intrade}_{t+1} = \alpha + \beta\text{Intrade}_t + \gamma\text{HSX}_t + noise$	0.99 (0.01)*	0.01 (0.01)
$\text{HSX}_{t+1} = \alpha + \beta HSX_t + \gamma\text{Betfair}_t + noise$	1.00 (0.00)*	0.02 (0.00)*
$\text{HSX}_{t+1} = \alpha + \beta HSX_t + \gamma\text{Intrade}_t + noise$	0.97 (0.01)*	0.04 (0.00)*

We obtain translated prices using the same methodology as for the raw polls, again assuming a linearly varying coefficient $\beta(t) = \beta_0 + \beta_1 t$. Similarly to polling we reduce the number of outcomes in each category to five. In the Sound Editing category, which had two winners, we assume that the actual winner is *Zero Dark Thirty*.[9]

The fitted translation function for Betfair has the form:

$$\beta(t) = 1.515 + 0.02t$$

where t ranges from $t = -40$ at the beginning of our evaluation period (40 days before Oscars) through $t = 0$ on the day of Oscars. The cross-validated standard errors for the coefficients are ± 0.015 (for the intercept β_0) and $\pm 3 \times 10^{-4}$ (for the slope β_1).

The function $\beta(t)$ is increasing over time, which means that as the forecasted event approaches, the raw prices increasingly underestimates the probability of the victory of the top candidate. This is likely due to an increasing favorite-longshot bias as more prices reach the extremes and transaction costs and risk-loving behavior prevent them from reaching the underlying subjective probabilities of the traders.

[9] We chose this approach due to the implementation convenience. A more correct approach would be to use the weighted log likelihood with the two winners corresponding to separate observations with weights 0.5. Since this is only one of 24 categories, the effect of this choice is negligible. *Zero Dark Thirty* was the leading nominee in the prediction markets throughout the evaluation period. The other winner, *Skyfall*, was 2nd for 27 days and 3rd for 14 days in the evaluation period.

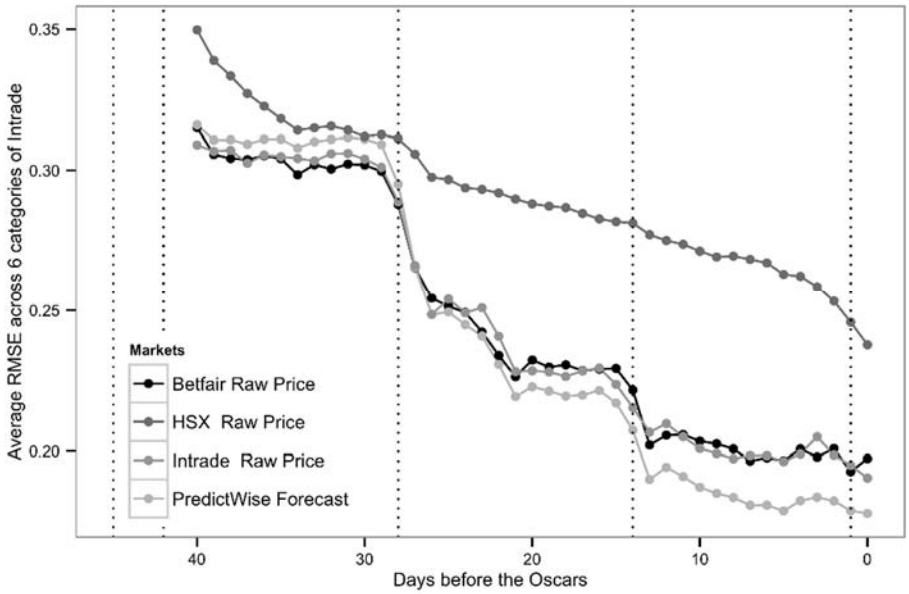

Figure 4. ***RMSE of naive (raw price) forecast from prediction market and the PredictWise data for six common categories*****. Average error across the six categories covered by all prediction markets for the 2013 Oscars. Prior awards shows are represented by vertical dotted lines, from left to right: Critics' Choice, Golden Globes, Guild, BAFTA, and Spirits.**

In order to judge accuracy of any measurable outcome, it is useful to examine both the error and calibration. Figure 5 shows the comparative analysis of error across Betfair raw prices, translated prices, and PredictWise. The figure shows the benefit of transforming raw prices as both translated Betfair and PredictWise have lower errors than the raw prices. The PredictWise model performs well in the beginning of the evaluation period, but the translated Betfair pulls ahead later in the evaluation period. While we do not show it in this figure, translated Intrade is very similar to translated Betfair in common categories and translated HSX has a much higher error over common categories. The general result of translated forecasts outperforming raw prices also holds in 2014 (not shown). In Appendix B, Figure 11 demonstrates that these outcomes are also well calibrated; if the forecast calls for 20% likelihood of an outcome, it happens about 20% of the time. While the raw prices are fairly well calibrated, translation further improves calibration. Note that the translation was derived from 2013 data, so Figure 11 should be interpreted with some caution. The calibration of the 2014 data was unaffected by the translation (not shown).

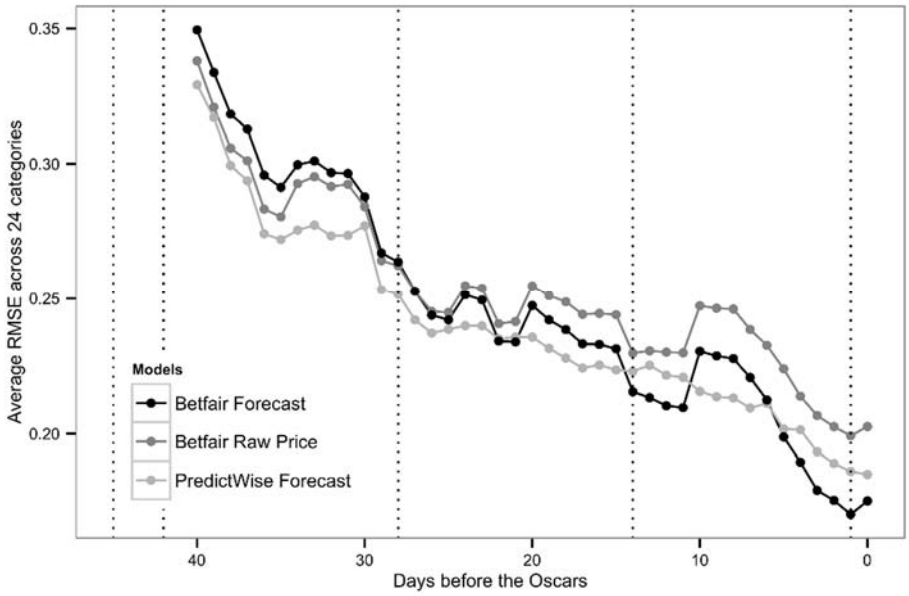

Figure 5. ***RMSE of raw prices and translated prices from prediction markets for all 24 categories.*** **Error across all 24 categories using raw Betfair, translated Betfair, and PredictWise for the 2013 Oscars. Prior awards shows are represented by vertical dotted lines, from left to right: Critics' Choice, Golden Globes, Guild, BAFTA, and Spirits.**

2.6 EXPERTS

Domain experts produce forecasts that are easily accessible by many stakeholders, so it is important to understand their advantages and disadvantages. We divide the experts in this domain into two groups. The first group contains movie pundits who use their critical skills and domain-specific expertise to discuss the outcomes. The second group contains the experts of the age of "big data", who use statistical models to construct forecasts from available quantifiable data.

Numerous pundits that fall into the first group publish their critical reviews and likely winners on the web. Their decisions are likely based on personal hunches or some undefined fundamental model. We referred to metacritic.com which presented aggregates (simple averages) of probabilistic predictions from 40 different pundits and entertainment writers across all categories. These aggregates were released on February 21, 2013, 3 days before the Oscars. An obvious drawback with this data is the lack of timeliness.

Table 3. *RMSE of expert forecasts.* Average error across all categories forecasted by experts for the 2013 Oscars. The three expert forecasts are for different sets of categories. We also include the RMSE of the Betfair forecasts for the same categories taken at the same point in time.

Experts	Days before the Oscars	Categories	Average RMSE	
			Experts	Betfair Forecast
Avg. of Oscar Pundits	3	24	0.20	0.18
Nate Silver	3	6	0.26	0.18
Ben Zauzmer	8-9	21	0.25	0.20

The statistical experts make data-based predictions that are similar to fundamental models. For example, Nate Silver of the New York Times presented his forecasts for three years—2009, 2011, and 2013,[10] while also publishing his data and models, which is highly unusual for experts. Initially in 2009, Silver used regression models over all relevant variables. Then in 2011, he simplified his models, keeping only logical variables which had good predictive power. In 2013, he focused only on the other awards shows as the predictors; this choice is also justified by our findings on fundamental data. The fact that he used a different model every year highlights two key concerns about expert forecasts. First, there is a risk that late-season changes in the data and methods can lead to a look-ahead bias (i.e., knowing the current expected outcome can affect choices for data and methods, yielding a model that produces results that herd with others). Second, similarly to many non-academic modelers, there is little protection from over-reliance on past results, which may not always generalize to future outcomes (i.e., within-sample models that do not work out-of-sample). Apart from ability to generalize to future, there are two additional concerns: timeliness and cost. In particular, experts tend to produce forecasts at a single point in time, and their restricted attention to a subset of categories can be seen as an evidence of a higher cost of forecasting. In addition to Nate Silver, we also consider forecasts of Ben Zauzmer, the next most cited expert.

Table 3 presents the average root mean square error for 2013 forecasts by three different experts data sources: (1) Oscar pundits' aggregate score, (2) Nate Silver and (3) Ben Zauzmer. Both Nate Silver's and Ben Zauzmer's predictions are based on their own fundamental models. For reference, we compare the accuracy of experts with the Betfair forecast. Since the Oscar pundits aggregate covers all categories and has fairly high accuracy, we continue to report it in our comparisons.

[10] http://nymag.com/movies/features/54335/
http://carpetbagger.blogs.nytimes.com/2011/02/24/4-rules-to-win-your-oscar-pool
http://fivethirtyeight.blogs.nytimes.com/2013/02/22/oscar-predictions-election-style/

2.7 COMPARISON OF ALL METHODS

Fundamentals, prediction markets, and experts all have examples of forecasts in all 24 categories for 2013. Figure 6 compares their average RMSE over all 24 categories. Prediction markets are significantly more accurate than all other models. The fundamental model catches up a little when there is a burst of information, but the gap remains large. The main reason for inferior accuracy of the fundamental model is a poor performance in the low information categories. There are 9 categories without any or very few corresponding awards and their average error right before the Oscars is 0.40, compared with the average error of 0.25 in the remaining 15 categories.

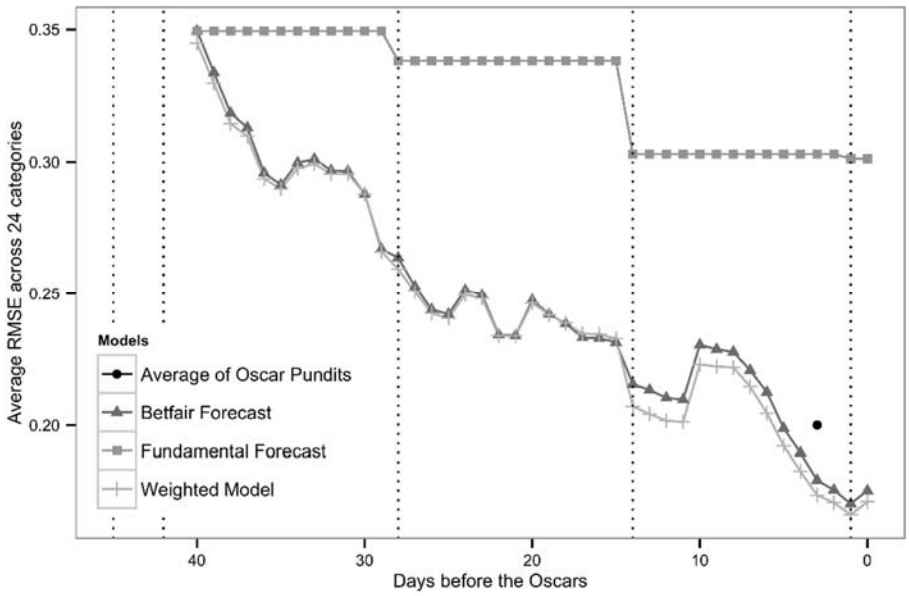

Figure 6. Average RMSE of Oscar Pundits, translated Betfair prices, fundamental model and combined weighted model for all 24 categories of the 2013 Oscars. Prior awards shows are represented by vertical dotted lines, from left to right: Critics' Choice, Golden Globes, Guild, BAFTA, and Spirits.

Figure 7 shows the errors of all forecasting methodologies across the nine common categories. First, fundamental model is extremely accurate at times of high information flow, and essentially matches the performance of prediction markets at points when the results of the Guild Awards and BAFTA are announced, but falls behind between the awards shows. Second, the polling does extremely well, especially at times of high engagement, such as 20 days before the Oscar night during a big social media push for self-selected respondents, or one day before the Oscar night during an increased push of the polls to random respondents.

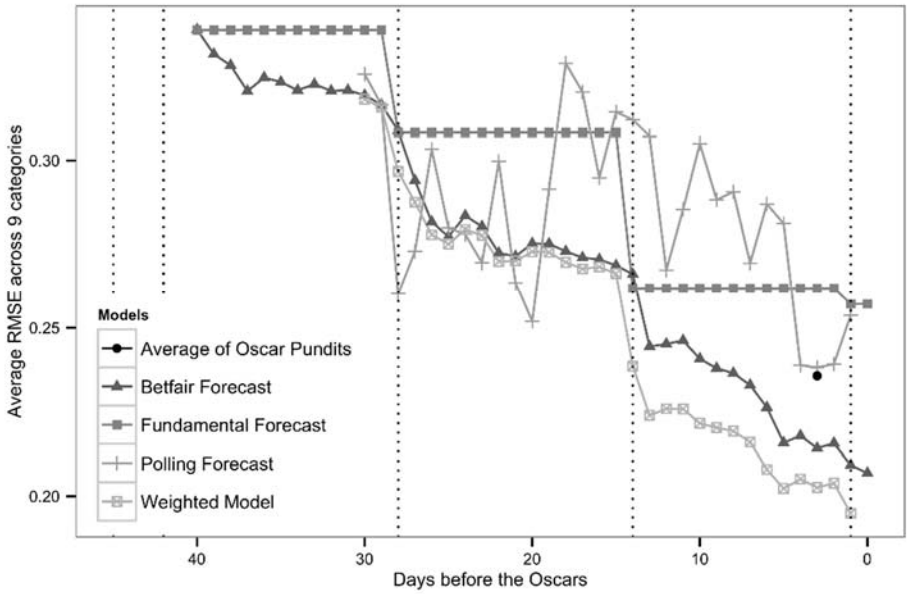

Figure 7. Average RMSE of Oscar Pundits, translated Betfair prices, fundamental model, translated polls (both types of respondents) and combined weighted model for the 2013 Oscars. Average error across the nine common categories covered by all the models. Prior awards shows are represented by vertical dotted lines, from left to right: Critics' Choice, Golden Globes, Guild, BAFTA, and Spirits.

2.8 COMBINING FORECASTS

In our final comparison we study the value of combining the forecasts. We test two different methods: simple average of forecasted probabilities and weighted average of forecasted log odds, referred to as the *weighted model.* They are constructed as follows. We consider combining either 3 models (fundamental, polling and Betfair, across 9 common categories) or 2 models (fundamental and Betfair, across all 24 categories). If the three models forecast probabilities p_1, p_2 and p_3 of a given nominee winning a given category, the simple average is just $p = (p_1 + p_2 + p_3)/3$. The weighted model is calculated in two stages, first we obtain $\hat{p}$ as:

$$\log\frac{\hat{p}}{1-\hat{p}} = \beta_1 \log\frac{p_1}{1-p_1} + \beta_2 \log\frac{p_2}{1-p_2} + \beta_3 \log\frac{p_3}{1-p_3}$$

and then rescale the values $\hat{p}$ for the nominees in the same category to sum to one. The coefficients β_1, β_2 and β_3 are fitted by logistic regression on 2013 data across all time points, all categories (either 9 or 24) and five nominees in

each category (we reduce the number of nominees as when fitting the translation). This is similar to the approach advocated by Rothschild (2015).

The simple average, while generally performing well in a variety of settings (Clemen and Winkler 1986), does not appear to work in this domain. The reason is apparent when we examine the coefficients of the weighted model. For 9 categories, we obtain weights $\beta_{\text{fund}} = 0.57$, $\beta_{\text{poll}} = -0.04$ and $\beta_{\text{Betfair}} = 0.63$. For 24 categories, we obtain weights $\beta_{\text{fund}} = 0.41$ and $\beta_{\text{Betfair}} = 0.82$. In both cases, the bulk of the meaningful prediction is on the prediction-market forecast. As shown in Figures 6 and 7, the weighted model runs extremely close to the prediction market forecast. We do not plot the simple average whose performance is inferior to the weighted model.

2.9 OUT-OF-SAMPLE EVALUATION ON THE 2014 OSCARS

While the fundamental forecasts are fully out-of-sample for 2013, the translated poll and prediction-market forecasts are not. In this section, we use those same models, completely out-of-sample, on the 2014 Oscars data.[11]

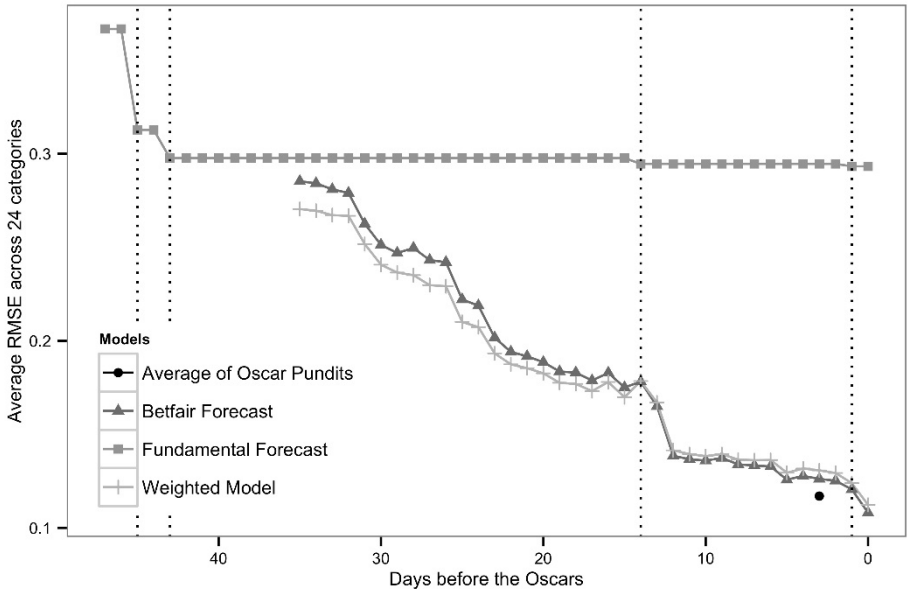

Figure 8. Average RMSE of Oscar Pundits, translated Betfair prices, fundamental model and combined weighted model for all 24 categories of the 2014 Oscars (out-of-sample). Prior awards shows are represented by vertical dotted lines, from left to right: Critics' Choice, Guild, BAFTA, and Spirits.

In Figures 8 and 9 we provide this fully out-of-sample comparison. For simplicity, we use the Civic Science polling, just as in 2013, and Betfair,

[11] The prediction-market model was published online in real time.

which was the dominant prediction-market model in 2013. In 2014, the prediction-market forecast dominates both the fundamental and polling forecasts.

Fundamental forecasts never approach the prediction-market forecasts in Figure 8, unlike their within-sample performance on the 2013 data. This is an indication that the year 2014 was more idiosyncratic than previous years. The earlier awards shows did not correlate as cleanly with the Oscars in 2014 as they did in 2013, compared with the wisdom of the prediction-market crowd. Figure 9 shows that despite the promise in 2013, the polling did not do as well in 2014. Again, this has a bit to do with the idiosyncratic nature of 2014 where popular movies did not do as well as they did in 2013, leading the polling public astray, but not the carefully incentivized prediction-market crowd.

3 DISCUSSION

This paper provides new insights into the relative value of different forms of data for creating forecasts. The academic literature tends to compare forecasts along one dimension, accuracy, at a single point in time, but that is not adequate in any practical sense. In particular, accuracy should not just mean a small error right before an event, but also robustness and calibration at other points in time. The timeliness is important for stakeholders and researchers alike, meaning both early forecasts and frequent forecasts. For the Oscars that means debuting the forecasts at the nominations and updating them continuously until the broadcast of the awards show. In addition to accuracy and timeliness, cost-effectiveness or the ability to scale the forecasts to new questions and domains is central to actual creation of the forecast. For the Oscars, this means the ability to cover all 24 categories and all nominees.

Fundamental data is expensive to translate into a forecast and that cost is not commensurate with the timeliness and accuracy of the forecast. In order to make forecasts in all 24 categories, there is an extraordinary cost in both data collection (which is category specific) and modeling (which is specific to categories and data availability at any given time). The timeliness of the forecast is limited by the availability of the prior awards show data. On one hand, since the forecast needs to wait until the first awards shows occur, it lacks most of its information at the nomination day. On the other hand, in 2013, the awards show data was mostly complete by 13 days before the Oscars, after the BAFTA awards, so the fundamental model did not continue to update towards the Oscar night. Thus, unless the model is evaluated at its most opportune moment, it is relatively not accurate throughout the season.

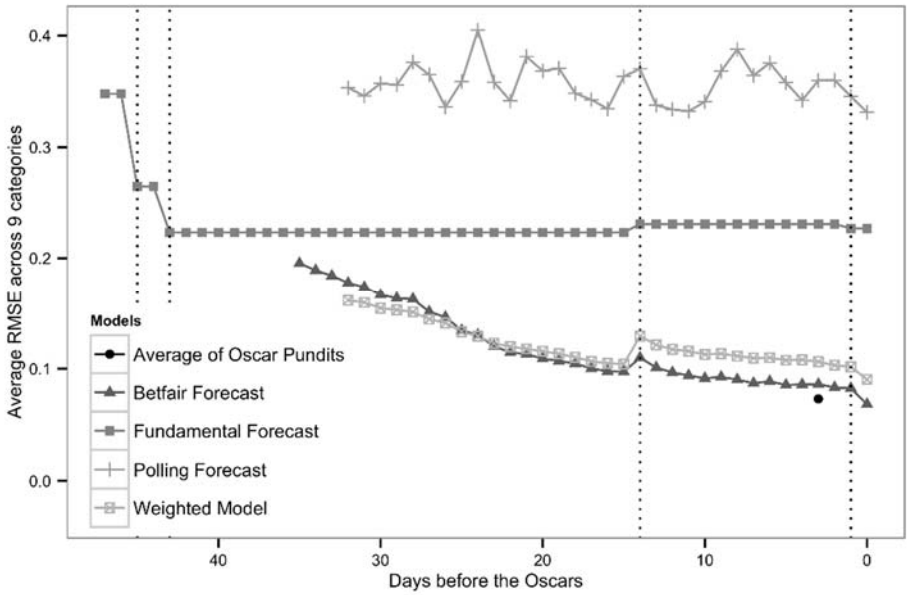

Figure 9. Average RMSE of Oscar Pundits, translated Betfair prices, fundamental model, translated polls and combined weighted model for nine common categories of the 2014 Oscars (out-of-sample). Average error across the nine common categories covered by all the models. Prior awards shows are represented by vertical dotted lines, from left to right: Critics' Choice, Guild, BAFTA, and Spirits.

Polling data shows a lot of promise in this domain. The self-selected responses translate into accurate forecasts 20 full days before the Oscars and the translated random responses are increasingly accurate as the Oscars approach. We only have data from the nine biggest categories, but it is possible to ask about all 24 categories. It is likely that the more obscure the category, the larger the divide between random and self-selected respondents; the random respondents will exhibit less accuracy as the categories become more obscure, while the self-selected respondents are expected to be better informed. The expectation questions used in the polling are similar to implicit questions in prediction markets, so it is not surprising the questions yield accurate forecasts with informed users. However, polls do not provide the possibility of monetary rewards that would keep respondents engaged on a continuous basis.

Prediction-market forecasts excel in all aspects: accuracy, timeliness and cost-effectiveness. There is minimal marginal cost to creating forecasts for all 24 categories as there is a low marginal cost of having all categories after a market exists. It is unfortunate that Intrade and HSX did not include all categories, because Betfair demonstrated very accurate results, even with the low liquidity for the more obscure categories. Prediction markets move in real

time as events unfold. And, they are extremely accurate: low errors and impressive calibration.

Our final data source, expert forecasts, are less timely than other options and their accuracy is no better than that of prediction markets. Also, the cost is a factor which limits the coverage of the questions.

The observations and methods developed in this paper should extend beyond the domain of the Oscars. Our critiques of the various data sources and their transformations into forecasts confirm and expand on the body of literature noted in the introduction, which spans numerous domains.

4 REFERENCES

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., et al., 2008. The promise of prediction markets. *Science* 320 (5878), 877.

Clemen, R. T., Winkler, R. L., 1986. Combining economic forecasts. *Journal of Business & Economic Statistics* 4 (1), 39-46.

Diebold, F. X., Mariano, R. S., 2002. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 20 (1).

Erikson, R. S., Wlezien, C., 2008. Are political markets really superior to polls as election predictors? *Public Opinion Quarterly* 72 (2), 190-215.

Fair, R., 2011. Predicting presidential elections and other things. Stanford University Press.

Ghitza, Y., Gelman, A., 2013. Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. American Journal of Political Science.

Goel, S., Reeves, D. M., Watts, D. J., Pennock, D. M., 2010. Prediction without markets. In: Proceedings of the 11th ACM conference on Electronic commerce. ACM, pp. 357-366.

Granberg, D., Brent, E., 1983. When prophecy bends: The preference-expectation link in U.S. presidential elections. *Journal of Personality and Social Psychology* 45 (3), 477-491.

Granger, C. W., Ramanathan, R., 1984. Improved methods of combining forecasts. *Journal of Forecasting* 3 (2), 197-204.

Guedj, O., Bouchaud, J.-P., 2005. Experts' earning forecasts: Bias, herding and gossamer information. *International Journal of Theoretical and Applied Finance* 8 (7), 933-946.

Harvey, D. S., Leybourne, S. J., Newbold, P., 1998. Tests for forecast encompassing. *Journal of Business & Economic Statistics* 16 (2), 254-259.

Hong, H., Kubik, J. D., Solomon, A., 2000. Security analysts' career concerns and herding of earnings forecasts. *The Rand Journal of Economics*, 121-144.

Hummel, P., Rothschild, D., 2014. Fundamental models for forecasting elections at the state level. *Electoral Studies* 35, 123-139.

Lock, K., Gelman, A., 2010. Bayesian combination of state polls and election forecasts. Political Analysis 18 (3), 337-348.

Manski, C. F., 2006. Interpreting the predictions of prediction markets. *Economics Letters* 91 (3), 425-429.

Pennock, D. M., Lawrence, S., Giles, C. L., Nielsen, F. A., et al., 2001. The real power of artificial markets. *Science* 291 (5506), 987-988.

Rothschild, D., 2009. Forecasting elections comparing prediction markets, polls, and their biases. *Public Opinion Quarterly* 73 (5), 895-916.
Rothschild, D., 2015. Combining forecasts for elections: Accurate, relevant, and timely. *International Journal of Forecasting* 31, 952-964.
Rothschild, D., Wolfers, J., 2011. Forecasting elections: Voter intentions versus expectations. Available at SSRN: http://ssrn.com/abstract=1884644.
Squire, P., 1988. Why the 1936 Literary Digest poll failed. Public Opinion Quarterly 52 (1), 125-133.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.* Series B (Methodological), 267-288.
Wang, W., Rothschild, D., Goel, S., Gelman, A., 2015. Forecasting elections with non-representative polls. *International Journal of Forecasting* 31, 980-991.
Wolfers, J., Zitzewitz, E., 2004. Prediction markets. Tech. rep., National Bureau of Economic Research.
Wolfers, J., Zitzewitz, E., 2006. Interpreting prediction market prices as probabilities. Tech. rep., National Bureau of Economic Research.

5 APPENDIX A. FINAL FUNDAMENTAL MODEL FOR ALL CATEGORIES

Table 4. ***Coefficients in the final fundamental model.*** **Standard errors provided in parentheses. Dashes indicate that the coefficient and its standard deviation are zero.**

Variables	Picture	Directing	Actor	Sup Actor	Actress	Sup Actress
Constant Gross/Screen	------	------	------	------	------	------
Slope Gross/Screen	------	------	------	------	------	------
Week Wide	0 (0.051)	0 (0.035)	0 (0.058)	0.005 (0.048)	0 (0.027)	0 (0.047)
Gross/Screens Wide	------	------	------	------	------	------
Release Date	0.002 (0.002)	------	0.001 (0.002)	------	0 (0.001)	0 (0.001)
Popular Rating	0 (0.014)	0 (0.015)	0.021 (0.028)	0 (0.006)	0 (0.008)	-0.014 (0.016)
Critical Rating	-0.006 (0.017)	------	0 (0.005)	0 (0.002)	0 (0.008)	0 (0.006)
Oscar Overall Nom	0.243 (0.109)	0.070 (0.060)	0 (0.047)	------	0 (0.019)	0.105 (0.081)
Critics Overall Nom	0 (0.020)	0 (0.003)	0 (0.007)	0 (0.003)	0 (0.030)	-0.043 (0.052)
Critics Overall Win	0 (0.054)	0 (0.004)	0 (0.069)	0 (0.108)	0 (0.093)	0 (0.040)
Critics Category Nom	0 (0.383)	0.320 (0.397)	0 (0.229)	0 (0.110)	0 (0.084)	0 (0.203)
Critics Category Win	1.127 (0.826)	2.824 (0.731)	1.012 (0.672)	0 (0.313)	0 (0.322)	0 (0.301)
GG Overall Nom	0 (0.061)	0 (0.023)	0 (0.032)	0 (0.003)	0 (0.021)	-0.055 (0.065)
GG Overall Win	0.232 (0.202)	0 (0.024)	0 (0.022)	0 (0.014)	0 (0.194)	0 (0.076)
GG Category Nom	0 (0.199)	0 (0.003)	0 (0.109)	0 (0.059)	0 (0.035)	0 (0.159)
GG Category Win	0 (0.148)	0.390 (0.533)	0.390 (0.463)	1.576 (0.634)	1.355 (0.565)	1.686 (0.607)
Guild Overall Nom	0.209 (0.190)	0 (0.051)	0.183 (0.151)	0 (0.028)	0 (0.033)	-0.121 (0.134)
Guild Overall Win	0.376 (0.301)	0.023 (0.158)	0.346 (0.353)	0 (0.121)	0 (0.191)	0 (0.080)
Guild Category Nom	------	------	0 (0.118)	0 (0.098)	0 (0.028)	0 (0.136)
Guild Category Win	------	------	2.563 (0.785)	1.406 (0.644)	2.134 (0.695)	0.225 (0.475)
BAFTA Overall Nom	0 (0.019)	0 (0.001)	0 (0.017)	0 (0.001)	0 (0.010)	0 (0.014)
BAFTA Overall Win	0.357 (0.189)	0.048 (0.098)	0.015 (0.057)	0 (0.029)	0 (0.047)	0 (0.052)
BAFTA Category Nom	0 (0.255)	0 (0.064)	0 (0.211)	0 (0.114)	0 (0.193)	0 (0.199)
BAFTA Category Win	0 (0.363)	0 (0.245)	0 (0.314)	0 (0.332)	0.906 (0.601)	2.376 (0.759)
Spirit Overall Nom	-0.040 (0.078)	0 (0.005)	0 (0.013)	0 (0.016)	0 (0.005)	-0.016 (0.061)
Spirit Overall Win	0 (0.045)	------	0 (0.033)	0 (0.043)	0.166 (0.143)	0 (0.080)
Spirit Category Nom	------	------	0 (0.047)	0 (0.181)	0 (0.038)	0 (0.404)
Spirit Category Win	------	------	0 (0.047)	0 (0.268)	0 (0.151)	0 (0.293)
Constant	-5.265 (2.263)	-3.106 (1.215)	-4.854 (2.262)	-2.216 (0.617)	-2.981 (1.027)	-1.439 (1.473)

Table 5. ***Coefficients in the final fundamental model.*** **Standard errors provided in parentheses. Dashes indicate that the coefficient and its standard deviation are zero.**

Variables	Adapted Screenplay	Original Screenplay	Song	Score	Sound Mixing	Sound Editing
Constant Gross/Screen	------	------	------	------	------	------
Slope Gross/Screen	------	------	------	------	------	------
Week Wide	0 (0.037)	0 (0.030)	0 (0.062)	0.153 (0.114)	0.002 (0.114)	0 (0.106)
Gross/Screens Wide	------	------	------	------	------	------
Release Date	------	0 (0.001)	0.001 (0.002)	0 (0.001)	0 (0.001)	0 (0.001)
Popular Rating	0 (0.012)	0.045 (0.028)	0 (0.005)	-0.003 (0.014)	0 (0.011)	0 (0.006)
Critical Rating	0 (0.009)	0 (0.007)	0 (0.005)	0 (0.012)	0.001 (0.006)	0 (0.007)
Oscar Overall Nom	0 (0.024)	0.021 (0.039)	0 (0.017)	0.029 (0.049)	0.003 (0.038)	0.102 (0.075)
Critics Overall Nom	0 (0.009)	0 (0.040)	0 (0.008)	0 (0.026)	0 (0.020)	0 (0.018)
Critics Overall Win	0.324 (0.206)	0 (0.045)	0 (0.058)	0 (0.044)	0.149 (0.150)	0 (0.044)
Critics Category Nom	0 (0.116)	0.322 (0.482)	0 (0.120)	0 (0.109)	0 (0.311)	------
Critics Category Win	0 (0.299)	1.511 (0.786)	0.973 (0.644)	0.060 (0.450)	0 (0.687)	------
GG Overall Nom	0 (0.022)	0 (0.032)	-0.037 (0.052)	0 (0.037)	0 (0.025)	0 (0.084)
GG Overall Win	0.480 (0.217)	-0.151 (0.197)	0 (0.017)	0.359 (0.228)	0 (0.080)	0 (0.111)
GG Category Nom	------	0 (0.356)	-0.015 (0.179)	0 (0.261)	------	------
GG Category Win	------	2.083 (0.926)	1.215 (0.709)	2.206 (0.830)	------	------
Guild Overall Nom	0.004 (0.088)	0.007 (0.096)	-0.029 (0.108)	0 (0.096)	0.055 (0.101)	0 (0.052)
Guild Overall Win	0 (0.206)	0.327 (0.379)	0 (0.062)	0 (0.126)	0 (0.187)	0 (0.148)
Guild Category Nom	------	------	------	------	------	------
Guild Category Win	------	------	------	------	------	------
BAFTA Overall Nom	0 (0.004)	0 (0.007)	0 (0.004)	0.016 (0.034)	0 (0.009)	0.006 (0.054)
BAFTA Overall Win	0 (0.091)	0 (0.087)	0 (0.043)	0 (0.044)	0.089 (0.149)	0.144 (0.231)
BAFTA Category Nom	0 (0.106)	0.776 (0.416)	0.255 (0.561)	------	1.202 (0.506)	------
BAFTA Category Win	0 (0.162)	1.857 (0.691)	0 (0.273)	------	1.354 (0.787)	------
Spirit Overall Nom	0 (0.018)	0 (0.038)	0.018 (0.146)	0 (0.060)	0 (0.143)	0 (0.009)
Spirit Overall Win	0 (0.071)	0.100 (0.184)	0 (0.179)	0 (0.204)	0 (0.365)	------
Spirit Category Nom	------	0 (0.136)	------	------	------	------
Spirit Category Win	------	1.062 (0.769)	------	------	------	------
Constant	-2.253 (1.245)	-7.168 (2.479)	-1.951 (0.723)	-2.650 (1.700)	-3.246 (1.040)	-1.597 (0.840)

Table 6. ***Coefficients in the final fundamental model.*** **Standard errors provided in parentheses. Dashes indicate that the coefficient and its standard deviation are zero.**

Variables	Cinema-tography	Art Direction	Costume Design	Film Editing	Visual Effects	Makeup
Constant Gross/Screen	------	------	------	------	------	------
Slope Gross/Screen	------	------	------	------	------	------
Week Wide	-0.039 (0.072)	0 (0.015)	0 (0.047)	0 (0.013)	0 (0.111)	0 (0.030)
Gross/Screens Wide	------	------	------	------	------	------
Release Date	0 (0.002)	0 (0.003)	0 (0.001)	0 (0.001)	-0.001 (0.002)	0 (0.001)
Popular Rating	0 (0.005)	0 (0.005)	0 (0.010)	0 (0.002)	0 (0.011)	0 (0.005)
Critical Rating	0 (0.008)	0 (0.004)	-0.005 (0.011)	0 (0.003)	0 (0.008)	0 (0.004)
Oscar Overall Nom	0.269 (0.083)	0.058 (0.060)	0 (0.019)	0.001 (0.040)	0.137 (0.082)	0 (0.014)
Critics Overall Nom	0 (0.021)	0 (0.029)	0 (0.024)	0 (0.013)	0 (0.042)	0 (0.063)
Critics Overall Win	0.265 (0.179)	0 (0.030)	0.066 (0.163)	0 (0.074)	0.069 (0.118)	0 (0.084)
Critics Category Nom	-0.689 (0.597)	0 (0.214)	0 (0.067)	0 (0.242)	0 (0.319)	0 (0.439)
Critics Category Win	0 (0.614)	0 (0.516)	2.065 (0.908)	0 (0.944)	0 (0.178)	0 (0.108)
GG Overall Nom	0 (0.033)	0 (0.021)	0 (0.009)	0 (0.021)	0 (0.066)	0 (0.011)
GG Overall Win	0 (0.108)	0.451 (0.250)	0.460 (0.225)	0.069 (0.127)	0.381 (0.286)	0 (0.013)
GG Category Nom	------	------	------	------	------	------
GG Category Win	------	------	------	------	------	------
Guild Overall Nom	-0.510 (0.213)	0 (0.031)	0 (0.080)	0 (0.016)	0 (0.077)	0 (0.056)
Guild Overall Win	0 (0.215)	0 (0.125)	0 (0.092)	0 (0.061)	-0.732 (0.428)	0 (0.136)
Guild Category Nom	------	------	------	------	------	------
Guild Category Win	------	------	------	------	------	------
BAFTA Overall Nom	0 (0.022)	0 (0.030)	0 (0.035)	0 (0.001)	0.108 (0.070)	0 (0.006)
BAFTA Overall Win	0.222 (0.185)	0.082 (0.177)	0.036 (0.120)	0.064 (0.109)	0 (0.106)	0.159 (0.160)
BAFTA Category Nom	0.517 (0.418)	------	1.060 (0.545)	0.911 (0.407)	0.112 (0.391)	0 (0.297)
BAFTA Category Win	0.238 (0.546)	------	0.091 (0.481)	0 (0.316)	1.555 (0.678)	1.885 (0.654)
Spirit Overall Nom	-0.024 (0.071)	0 (0.046)	0 (0.189)	0 (0.021)	------	0 (0.051)
Spirit Overall Win	0 (0.108)	0 (0.122)	0 (0.064)	0 (0.019)	------	0 (0.043)
Spirit Category Nom	0 (0.213)	------	------	------	------	------
Spirit Category Win	0 (0.213)	------	------	------	------	------
Constant	-3.748 (0.929)	-2.249 (0.796)	-2.320 (1.131)	-2.625 (0.560)	-1.737 (1.135)	-1.378 (0.634)

Table 7. ***Coefficients in the final fundamental model.*** **Standard errors provided in parentheses. Dashes indicate that the coefficient and its standard deviation are zero.**

Variables	Animated Feature	Animated Short	Doc Feature	Doc Short	Foreign	Live Action Short
Constant Gross/Screen	------	------	------	------	------	------
Slope Gross/Screen	------	------	------	------	------	------
Week Wide	0 (0.004)	0 (0.027)	0 (0.010)	0 (0.063)	0 (0.025)	0 (0.013)
Gross/Screens Wide	------	------	------	------	------	------
Release Date	------	------	------	0 (0.001)	0 (0.001)	------
Popular Rating	0 (0.001)	0 (0.006)	0 (0.002)	0.014 (0.008)	0.005 (0.012)	0 (0.001)
Critical Rating	0 (0.004)	0 (0.003)	0 (0.001)	0 (0.002)	0.003 (0.004)	0 (0.001)
Oscar Overall Nom	0 (0.096)	------	0 (0.360)	------	0 (0.063)	------
Critics Overall Nom	0 (0.065)	0 (0.012)	0.286 (0.259)	------	0 (0.050)	------
Critics Overall Win	1.099 (0.541)	------	0.351 (0.333)	------	0.297 (0.341)	------
Critics Category Nom	0 (0.033)	------	0 (0.280)	------	0 (0.141)	------
Critics Category Win	2.017 (0.540)	------	0.351 (0.333)	------	0.297 (0.341)	------
GG Overall Nom	0 (0.119)	------	0.562 (0.399)	-0.244 (0.183)	0 (0.071)	0 (0.063)
GG Overall Win	0 (0.270)	------	0 (0.216)	-0.487 (0.366)	0.655 (0.369)	------
GG Category Nom	0 (0.039)	------	------	------	0.235 (0.418)	0 (0.188)
GG Category Win	0 (0.244)	------	------	------	0.537 (0.556)	------
Guild Overall Nom	------	------	0 (0.211)	------	0 (0.07)	------
Guild Overall Win	------	------	------	------	0.817 (0.659)	------
Guild Category Nom	------	------	------	------	------	------
Guild Category Win	------	------	------	------	------	------
BAFTA Overall Nom	0 (0.029)	0 (0.153)	0 (0.197)	------	0 (0.055)	------
BAFTA Overall Win	0.820 (0.622)	0 (0.132)	0 (0.337)	------	0 (0.039)	------
BAFTA Category Nom	0 (0.091)	0 (0.153)	------	------	0 (0.235)	------
BAFTA Category Win	1.282 (0.672)	0 (0.132)	------	------	0 (0.097)	------
Spirit Overall Nom	------	0 (0.003)	0 (0.035)	------	0 (0.018)	------
Spirit Overall Win	------	------	0 (0.183)	------	0 (0.204)	------
Spirit Category Nom	------	------	0 (0.035)	------	0 (0.123)	------
Spirit Category Win	------	------	0 (0.183)	------	0 (0.250)	------
Constant	-2.568 (0.500)	-0.902 (0.703)	-1.595 (0.524)	-2.002 (0.724)	-2.385 (0.900)	-1.304 (0.333)

6 APPENDIX B. ACCURACY AND CALIBRATION PLOTS

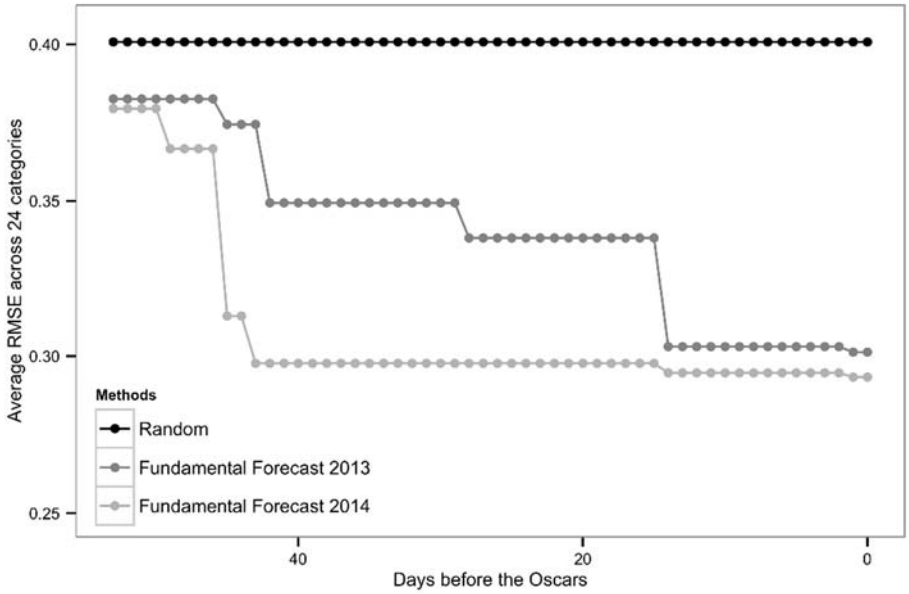

Figure 10. ***Average RMSE of fundamental model across all 24 categories*****. Out-of-sample error across all 24 categories for the 2013 and 2014 Oscars.**

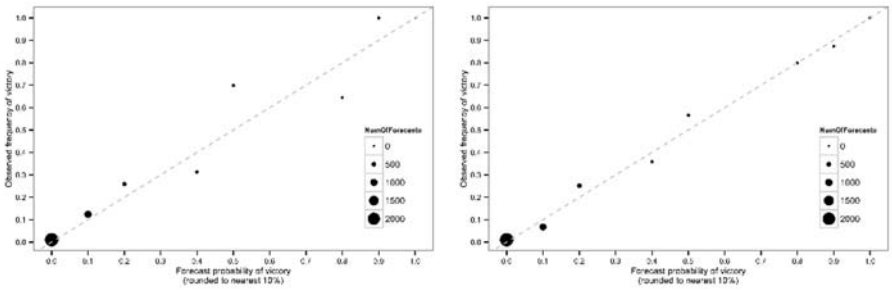

Figure 11. ***Calibration of Betfair Raw Prices (left) versus Betfair Translated Prices (right)*****. Across all 24 categories using translated Betfair prices once per day for 40 days for the 2013 Oscars.**