A comparison of forecasting methods: fundamentals, polling, prediction markets, and experts[☆]

Deepak Pathak[a], David Rothschild[b], Miroslav Dudík[b]

[a]*IIT Kanpur, India*
[b]*Microsoft Research, New York, USA*

Abstract

In this paper we dissect four different data types that can create forecasts for the results of the Oscars: fundamentals, polling, prediction markets, and experts. We start by creating the most accurate forecasts possible for each data type. We then compare and contrast how these different data types and their forecasts compete over four attributes: relevancy, accuracy, timeliness, and cost effectiveness. We argue that forecasts created with fundamental data are expensive to construct and show only limited adjustments over time, which constrains how accurate and timely the forecasts can be; the academic literature should not ignore cost-effectiveness in that it is critical to the practical application of the forecast research. However, fundamental data does provide valuable insights into the relationship between key indicators for nominated movies and outcomes; most of the power comes from the results in other awards shows, popular ratings are more important than critical ratings, and box office performance has little predictive power. Polling has the potential to create both accurate and timely forecasts, but it requires incentives for frequent responses by high-information users to stay timely, and proper transformation of raw polls into forecasts to be accurate. Third, prices in prediction markets are highly accurate forecasts, but can further benefit by a simple transformation, yielding the most accurate method in our comparison. Finally, experts create something similar to fundamental models, but are generally not comparatively accurate or timely. Overall, prediction markets lead our comparison across all attributes, but polling shows promise; we believe that the results of this study generalize to many domains.

Keywords: prediction, oscar, academy awards, regression, translation, probabilities

1. Introduction

The Oscars is an awards show for the movie industry which is watched live by millions of people. The most recent Oscar ceremony to the writing of this paper, held on February 23, 2013, was watched live by over 40 million viewers and ads for the next Oscar ceremony, on

[☆]The authors would like to thank David Pennock for all of his help with this project. We would like to thank Civic Science for providing us the polling data.

March 2, 2014, have already sold for between $1.8 million and $1.9 million per 30 seconds.[1] The 2013 show comprises the main set of outcomes discussed in this paper and that year there were awards in 24 categories that ranged from very visible work like several variations of best picture and best actor to behind the camera work like best director and best cinematography. Similar to any live event viewed by millions of people, people eagerly debate and bet on who will win the different outcomes. Further, studios that produce movies wage campaigns for their movies, because winning provides new interest in the movies.[2] Thus, there is a strong interest in and monetary incentive to forecast the winners accurately and see how these forecasts move over time. This paper examines several prominent data sources used to forecast the winners in the domain of the Oscars and then generalizes these results for many other domains.

The goal of any forecast, what makes an efficient forecast, is to be relevant, accurate, timely, and cost effective. The forecast is relevant if it provides the most useful information to the stakeholders. For the Oscars, this includes probabilities of victory for all nominees in all categories. 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 campaign season. 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. The forecast is cost effective if it is worth the investment to produce the forecast; extending beyond the Oscars, a cost effective forecast method is generally scalable for other questions and/or domains.

The four data type discussed in the paper include: fundamentals, polls, prediction markets, and 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. Polling, prediction markets, and experts are all elicited in order to answer a question. Polling is when researchers ask people what they want or what they think. Examples of this data include polls of peopled asked which nominee they think will win certain categories. Prediction markets are markets where people can wager on the outcome of an event. Examples of this data include markets where there are contracts worth $1 if a nominee wins the Oscar and worth $0 if the nominee does not win the Oscar. Finally, experts are people that state their opinion on the likelihood of different outcomes. Example of this data include movie columnists who state the probability that any given nominee will win selected categories.

Fundamental data can make relevant and accurate forecasts in some domains, but there are high construction costs to fundamental models and they are generally not timely. It

[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/`

is well understood that fundamental data can make accurate forecasts in many domains, including politics (Fair, 2011; Hummel and Rothschild, 2013) and movie box office returns (Goel et al., 2010). Further, some literature has explored fundamental data's place in the timeline of the events, showing how it is most meaningful when less idiosyncratic information is available, as fundamental models are not good at absorbing dispersed or idiosyncratic information; for example this is true early in the cycle in politics (Lock and Gelman, 2010; Rothschild, 2013), before the available data is supplemented by information from polls and prediction markets. These papers show that later in the cycle they fail to update with enough information or in a timely manner relative to other data types. For this paper fundamental data is comprised of: box office returns, screens, ratings, other awards, etc. that we use to construct statistical models. Creating fundamental models is generally expensive, because the correlation of fundamental data and outcomes is different for each outcome; for example, in the Oscars we need to construct a new model for each category. Further, also common to many domains, since some of the data is not available for the full timeframe, each category has multiple models to accommodate the data available at each timeframe.

Polling data can create relevant and accurate forecasts, but necessarily timely. There is a dense literature on random and representative polling creating accurate forecasts of upcoming events (Erikson and Wlezien, 2008). The literature is generally dismissive of the value of non-random and/or non-representative polling (Squire, 1988). But, there is a thin literature on the methods that could create value of non-representative polling (Ghitza and Gelman, 2013) and even less on the empirical results of that work (Wang et al., 2013). Yet, Rothschild and Wolfers (2011) demonstrates empirically how non-representative polling can benefit from asking more appropriate questions for the sample, such as the expectation question for aggregate forecast. The polling data we test in this paper is the expectation poll, in a selection of categories, administered to both self-selected and random respondents. Similar to standard polling, our data is going to be most accurate after it is collected and polls are rarely collecting consistent responses on a day by day basis; standard representative polling costs tens to hundreds of dollars per respondent to recruit and even non-representative polling requires some ad space or other active effort to recruit users.

Prediction markets are relevant, 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 even in movies (Pennock et al., 2001). We examined three different sets of prediction market data; two real-money markets and one play money market. The literature has shown that play money markets can work well, including the one used in this paper (Pennock et al., 2001).

Expert forecasts' accuracy can suffer due to the incentives of the forecaster and timeliness is not a key priority. It is well known in the literature that experts suffer from herding and over-reliance on within-sample models (Guedj and Bouchaud, 2005). Many experts are not always incentivized to provide the most accurate forecast; an expert may place their 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 updating their forecasts as new

information arrives, which would be very costly for one individual to do.

These four data types are common in forecasting situations and we expect these results to translate into other domains. First, fundamental data can be expensive to gather and it updates slowly, which constrains how accurate the forecasts can be at any given moment before an event; the academic literature should not ignore cost-effectiveness in that it is critical to the practical application of the forecast research. But, fundamental data does provide valuable insights into the relationship between key indicators for nominated movies and outcomes; most of the power comes from the other awards shows, popular ratings are more important than critical ratings, and box office performance has little predictive power. Second, polling data has the potential to create accurate forecasts, but it requires incentives for regular responses by high information users to stay updated and proper transformation of raw polls into forecasts to be accurate. Third, prediction markets prices are consistent forecasts, but properly transformed prices are the most efficient forecast. Fourth, experts create something similar to fundamental models, but are generally not comparatively accurate or timely.

The paper has three main findings. First we provide simple and reusable translation methodology for creating forecasts from raw fundamental, polling, and prediction market data. Second, we gain some interesting domain-specific insights from the fundamental data. Third, prediction market data creates the most relevant, accurate, and timely forecasts at scalable costs, but polling can provide significance under the right conditions.

2. Data, Estimation Strategy, and Results

2.1. Target Domain: The Oscars 2013

The main outcome variables for this paper are the 24 categories of Oscars awarded on February 24, 2013. For all but one category, Best Picture, there were five nominees per category, with the winner declared as the highest vote getter among the nominees. The Best Picture category had nine nominees in 2013 and this value will fluctuate between years. The voters, for both the nominees and the awards, are the approximately 6,000 members of the Academy of Motion Picture Arts and Sciences. 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

[3]The L.A. Times was able to contact what they believe is 88% of the Academy voters in 2012: `http://www.latimes.com/entertainment/la-et-movie-academy-methodology-html,0,4708801.htmlstory#axzz2jQn95BfG`

error (RMSE):

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

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:

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

2.3. Fundamental Data

Fundamental data describes a type of data that researchers do not necessarily collect to answer a forecasting question; the data exists for other reasons. The first step in creating a fundamental model is the data collection. This is costly in domains like the Oscars where each of the 24 categories has its own domain specific data to collect; further, not all data is available at all points during the forecasting period, with some awards data trickling in as Oscar night approaches. The outcome of other awards shows prove to be highly predictive, so it is critical for the accuracy of the forecast to collect and include them as they become available. The second step is creating the models that transform the raw data into forecasts. 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, how we derive the models, and then present the cross-time comparisons of accuracy.

Most of the fundamental data that people think of for movies, such as box office receipts, ratings, etc., is movie specific, so it provides little predictive power for categories that focus on very specific attributes of the movies. Thus, we supplement the data, wherever possible, with category specific data. The most prominent of this is the awards show data. Our initial pool of fundamental data includes the following data sources for all nominated movies from 1978 forward: name of person (if applicable), gross revenue and screens per week for the first eight weeks of release, release date, critical and popular rating, MPAA rating, genre, budget. Further we include awards shows as they unfold: Critic's Choice Award, Golden Globes, Guild Awards, British Oscars, and Spirits Awards. The box office data was collected from boxofficemojo.com website, corresponding records for other awards shows from IMDB, and rankings from Rotten Tomatoes.

Each piece of data we collected could be included in the models in many different ways, but simple tests of predictive power by variable allow us to quickly scale down the quantity of variables and determine the most efficient realization of the variables prior to running the main models. Despite collecting data from 1978 onward, we only use 1995 forward, because the earlier data actually makes the forecasts less accurate. We translate release date as the difference of days before 25^{th} February. We include critical and popular rating from Rotten Tomatoes as they are. We collected box office trend as gross revenue and screens for first eight weeks of release, which could be realized as many different variables; we include the box office data in several ways. We derive one variable by fitting a linear model over the

variation of revenue per screen over the first eight weeks and include that trend. We derive a second variable by the revenue from the movie's wide opening week defined as the week when movie is showing on more than 600 screens in the United States and Canada. The monetary values have been scaled to current currency value wherever required. Finally, we include both whether the Oscar nominee was nominated and won for both corresponding categories and overall in other awards shows namely Critic's Choice Award, Golden Globes, Guild Awards, British Oscars, and Spirits Awards. This is a non-trivial exercise as categories do not fit neatly between awards shows, so we match the categories by hand, and categories change over time. For one key example, in the Golden Globes there are division in terms of musical/comedy and drama, and we considered both to be equally relevant.

The variables are indexed by $k = 1, \ldots, d$. The years of data are 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. 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

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

where $\mathbf{X}_{tij}$ 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}. We drop the year 2013 from dataset to evaluate the model accuracy out-of-sample.

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

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

where the regularization coefficient λ is chosen by fivefold cross-validation.

With 144 models, six each for 24 categories, it is impossible to detail all coefficients, but here are some of the most interesting trends we notice about the predictive power of the variables across the 24 categories. For reference, we detail the coefficients for the final model for the all 24 categories in Tables A.5–A.8 in Appendix A and for all six models of Best Picture in Table 1. First, shown clearly in Table 1 with the 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 in Tables A.5–A.8;

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 works at a later point in time, until Model 6, which is the last model that can be determined just a few days before Oscar night. Standard errors provided in parentheses.
* The results of these awards shows were not declared 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)	0.002 (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)

there is very little predictive power to the other variables. Understanding that caveat, there are few interesting points. The critical ratings are not predictive, but the popular ratings are more predictive; this is illustrated best in the Original Screenplay category in Table A.6. The release date does matter, but in an unusual way. Our forecasts are on movies winning the Oscars, conditional on being nominated for an Oscar. While a movie is more likely to get a nomination if it opens later in the season, conditional on being nominated early released movies are a little more likely to win. Conditional on being nominated there is very little predictive power from the success in box office; this is shown in both Table 1 and in Tables A.5–A.8 across many categories.

Detailing the six models for the Best Picture shows the evolution of the forecasts through the awards season. Table 1 shows all of the coefficients from the six Best Picture category models; the models move in time from left to right with the first model applicable to the moment when the nominations are released while the last model is realized just days before Oscar night. The overall number of Oscar nominations is always valuable; this is not

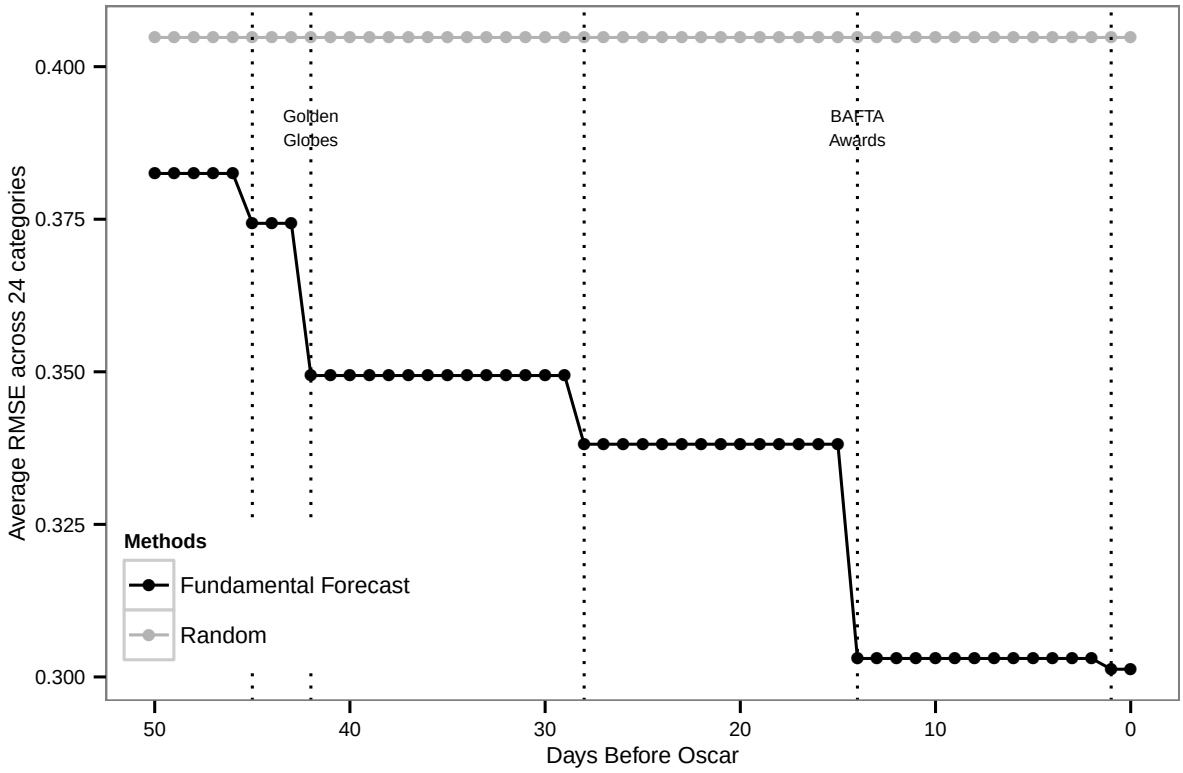

Figure 1: *Average RMSE of fundamental model across all 24 categories.* Out-of-sample error across all 24 categories for the 2013 Oscars. The awards shows are represented by vertical dotted lines, from left to right: Critic's Choice, Golden Globes, Guild, BAFTA, and Spirits.

surprising as the Oscar voters are likely to think well of the decision of the Oscar voters. Notice that awards show nominations are available from the beginning, but wins can only join after the awards show. After the awards show has occurred, the coefficient on wins can be upwards of a magnitude larger than the coefficient on nominations for a given awards show. All of the first four awards shows have similarly large coefficients, but the Spirit Awards, aimed towards independent films, provides little additional predictive power.

We judge the accuracy of our model by the average RMSE across all 24 categories. Figure 1 shows how our model compares with the random probabilistic model for which $p_{ij}^{random} = \frac{1}{m_i}$ 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: Critic's 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 the non-awards show 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 A.5–A.8.

2.4. Polling

Since the inception of representative polling in the 1930's, polling has been a central data type in forecasting upcoming events; in this paper we explore two different types of *non-representative* polling. Our polling data comes from a study by Civic Science.[4] Civic Science conducted the online public polling across nine categories for 40 days before the Oscars. They asked the online users who they expected to win the Oscars, as previous research indicates that the expectation questions are relatively informative questions when polling non-representative sample (Rothschild and Wolfers, 2011). Two separate user populations were surveyed. First, there is the "random" (but non-representative) population chosen randomly to take the Oscar questions 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, there is the "self-selected population", recruited to come to the Civic Science website and provide their expectation in up to seven specified categories. The self-selected respondents were recruited in two main ways. There was a massive social media push about 20 days prior to the Oscars and there was an ongoing link to the website for people who answered random polls powered by Civic Science and wanted to continue onto answer more questions. Note that both populations 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 cycle, mainly centered around the social media push to gather respondents, while random users bunch mainly late in the cycle, when it made sense for Civic Science to push Oscar questions to random respondents. There is no time period where both of them would show high activity. Second, the self-selected users were only provided seven questions, while the random users 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 time step considering only the polls on that day. Days will be indexed by t. Let $\tilde{p}_{ij}(t)$ denote the fraction of polls in i^{th} category supporting the j^{th} candidate; 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 Best Picture category, which has nine nominees, we keep the top

[4] Civic Science is contracted by third-party website to conduct their polls and subsequent data intelligence. They conduct polling for over 500 newspapers and blogs across the country.

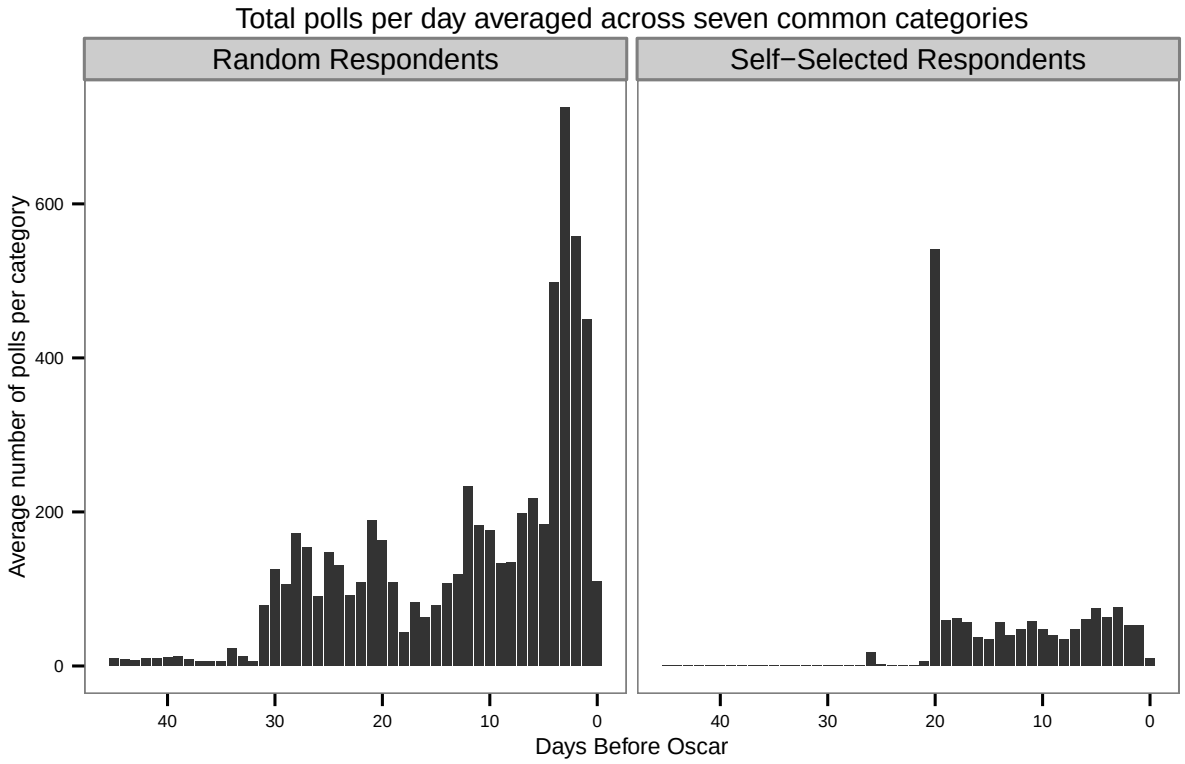

Figure 2: *Votes per day in the random and self-selected respondents for polling (across seven common categories of self-selected and random respondents).*

four polling nominees on a given day as separate outcomes and merge the bottom five into a single pseudo-nominee.[5]

Figure 3 compares the average RMSE across seven common categories at any time t, for both self-selected and randomly collected polls. We can see that forecasting error of the random users' expectation decreases with time; they are also adding more and more users per day over time and the available information is growing as the event nears. 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 site are not as knowledgeable as those that answered the poll on the first day. Recall that there was a massive social media push at 20 days out that targeted people with a potential interest in the Oscars and the main ongoing recruitment for the final 19 days was people trickling into the Civic Science website because they want to answer more questions in general. But, that error on the first day is very low; lower than any error produced by random polls even on the last day before the Oscars. This indicates that raw polls of a sufficient number of self-selected respondents can be substantially more accurate than those of random users. The self-selected respondents likely have more information per user,

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

providing less noisy responses, but, since they are more invested in the domain, they are also more likely to have strong preferences that may bias their stated expectations.[6] The random polling benefits from being having many more respondents, on average, demonstrating that enough polls among random users can still provide a meaningful result.

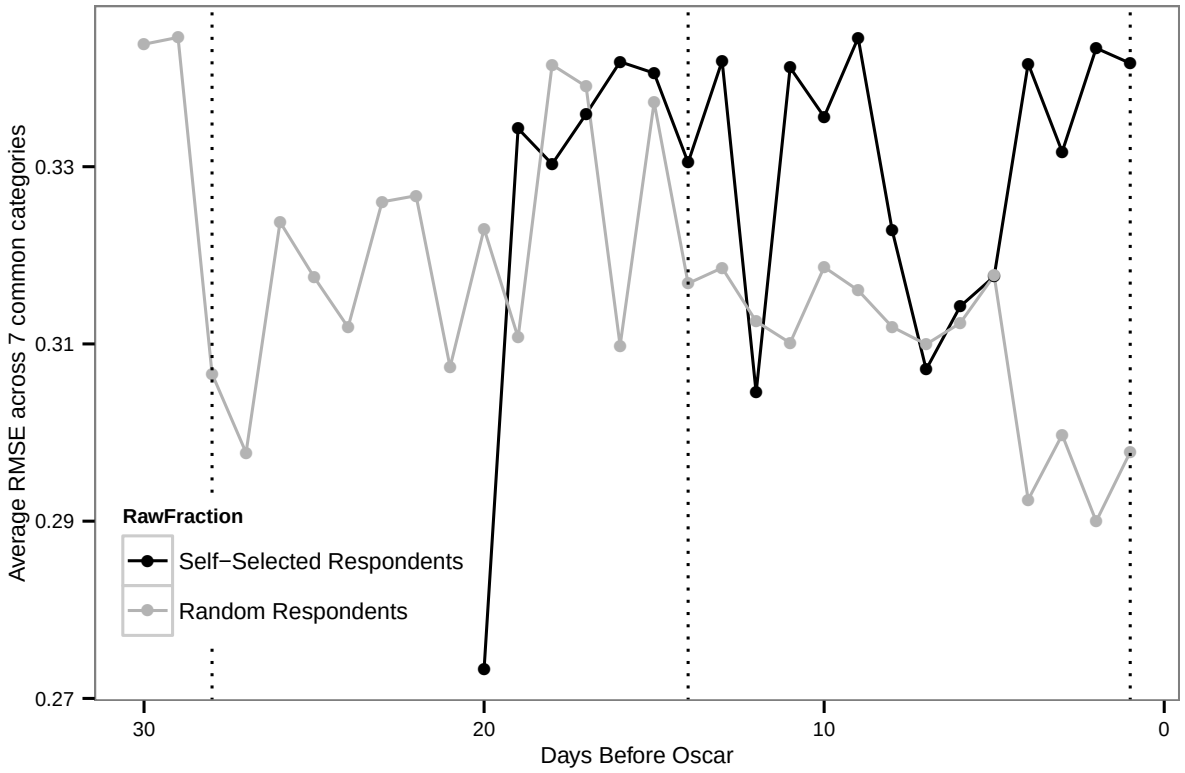

Figure 3: *RMSE of raw polls for all 7 common categories.* Error across all seven common categories of self-selected and random respondents for the 2013 Oscars. The awards shows are represented by vertical dotted lines, from left to right: Guild, BAFTA, and Spirits.

Now we consider constructing a more accurate forecast from the raw polling data. There is an accepted literature on the accuracy of forecasts created by applying a transformation to standard representative polls (Erikson and Wlezien, 2008). Our translation of raw poll results into a forecast 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 more probability than a nominee that polls at 25%.

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

[6]This is known as the wishful thinking bias in psychology.

category i (in our case it is always 5). Our *translated forecast* is of the form:

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

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 $\tilde{p}_{ij}(t) = \frac{n_{ij}(t)}{n_i(t)}$ and it differs by including a "pseudo vote" for each nominee (this is called 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:

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

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 constant function $\beta(t) = \beta_0$.

The resulting values of β_0 are reported in Table 2. In all considered datasets, the most accurate predictions are obtained by applying fairly large translation coefficients, in the approximate range 1.77 through 2.07. Thus, 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.

Table 2: *Translation coefficients for polling results.* Standard errors obtained by crossvalidation.

Dataset	**Coefficient $\beta_0 \pm$ std. error**
Self-selected respondents	2.03 ± 0.04
Random respondents	1.82 ± 0.05
All respondents	1.92 ± 0.08

Figure 4 shows that the translation of raw polls, using our methodology, gives significantly more accurate forecasts. Raw polls are plotted with [5-95]% confidence interval obtained using bootstrapping over the total polls over time for self-selected and randomly respondents. With a translation, the random respondents eventually create a similar error than the initial burst of self-selected respondents at 20 days before the Oscars, but not until last few days of the cycle. The translated results should be interpreted with caution, since the reported accuracy is within the same sample that we used to estimate the translation parameters β_0 and β_1, and we have no additional years to do an unbiased evaluation of our model. But, the results show the promise of our translation methodology.

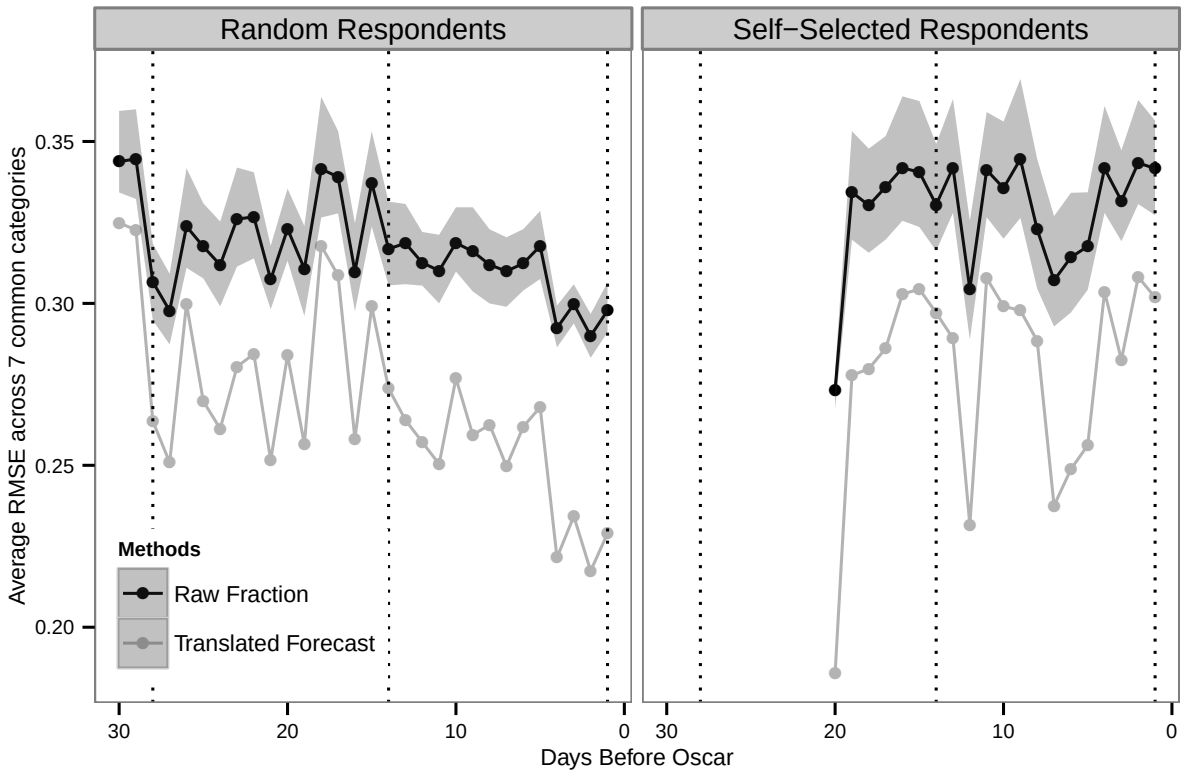

Figure 4: *RMSE of raw and translated polls across 7 common categories.* Error across all seven common categories of self-selected and random respondents for the 2013 Oscars. The awards shows are represented by vertical dotted lines, from left to right: Guild, BAFTA, and Spirits.

2.5. Prediction Markets

Prediction markets are markets where users can buy and sell contracts on upcoming events; the price of the security is highly suggestive of the probability of the outcome. For example there was 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%. We have data from three different prediction markets for this paper. We first compare how they influence each other. Then we examine the forecasting accuracy of raw prices and show how we can apply a necessary correction, analogous to polling. Finally, we illustrate how the forecasts compare.

The outcomes of Oscars are of widespread interest with huge impact on the film industry across all nations and public in general. The Oscars featured an active business primarily in three major prediction markets. This includes two real-money predictions markets—Betfair and Intrade—and one virtual-money prediction market—Hollywood Stock Exchange (HSX). While HSX is not real-money, they have a very loyal user base that treats the points they win and lose as meaningful. Other price we consider is PredictWise, which aggregated and

translated prices from Betfair and Intrade during the 2013 Oscar season.[7]

We will now describe the model used in PredictWise which showed their results live during the Oscar season. Let $\tilde{p}_{ij}^{Betfair}$ and $\tilde{p}_{ij}^{Intrade}$ denote the raw prices of the security for the j^{th} nominee in i^{th} category (at a particular instant of time). The PredictWise forecast was derived as

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

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, 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. Our data starts the Monday after announcement of the Oscar nominations and we record once per day at 23 : 00. The prediction markets were congruous and consistent with each other in terms of security price variation over time towards the Oscar day. Table 3 illustrates these correlations between the current prices on each exchange and the previous day's price on the other exchanges. In the first two rows we show that Betfair and Intrade have statistically significant and meaningful correlations with each other's price. The middle two rows show that, while technically statistically significant for Betfair, there is a very small correlation between HSX's earlier prices and Betfair's or Intrade's price. There does appear to be one interesting relationship in the last row, where Intrade's price from the previous day has statistical significance of a non-negligible size on HSX's current price.

Table 3: *Relationship between Betfair, Intrade, and HSX with lagged prediction market data of one day.* Statistically significant coefficients (at 1%) are denoted by $*$. Standard errors provided in parentheses.

Regression Formula	**Coefficient β**	**Coefficient γ**
$Betfair_{t+1} \sim \alpha + \beta Betfair_t + \gamma Intrade_t$	0.90 (0.028)*	0.10 (0.027)*
$Intrade_{t+1} \sim \alpha + \beta Intrade_t + \gamma Betfair_t$	0.80 (0.029)*	0.23 (0.029)*
$Betfair_{t+1} \sim \alpha + \beta Betfair_t + \gamma HSX_t$	0.99 (0.004)*	0.02 (0.005)*
$Intrade_{t+1} \sim \alpha + \beta Intrade_t + \gamma HSX_t$	0.99 (0.005)*	0.01 (0.007)
$HSX_{t+1} \sim \alpha + \beta HSX_t + \gamma Betfair_t$	1.00 (0.003)*	0.02 (0.002)*
$HSX_{t+1} \sim \alpha + \beta HSX_t + \gamma Intrade_t$	0.97 (0.005)*	0.04 (0.003)*

Interpreting raw security prices as forecasts has been widely studied (Manski, 2006; Wolfers and Zitzewitz, 2006). Figure 5 shows the average RMSE of raw prices across six common categories in Betfair, Intrade, HSX and PredictWise. First, all of the errors decrease

[7]Thus, all of PredictWise's aggregation and debiasing was done ex-ante and published publicly.

towards the Oscar show for all markets. Second, there are big drops in errors around the third and fourth awards shows. Third, the two real-money prediction markets are in virtual lock-step. That is good, it shows market efficiency. Fourth, the real-money markets have better calibration and consequently have much smaller errors than the virtual-money market. We use Betfair as our main prediction market data moving forward because it is almost the same as Intrade, but exists in all 24 categories, and it is much more accurate than HSX, both in raw prices and 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 consider translating the raw prices.

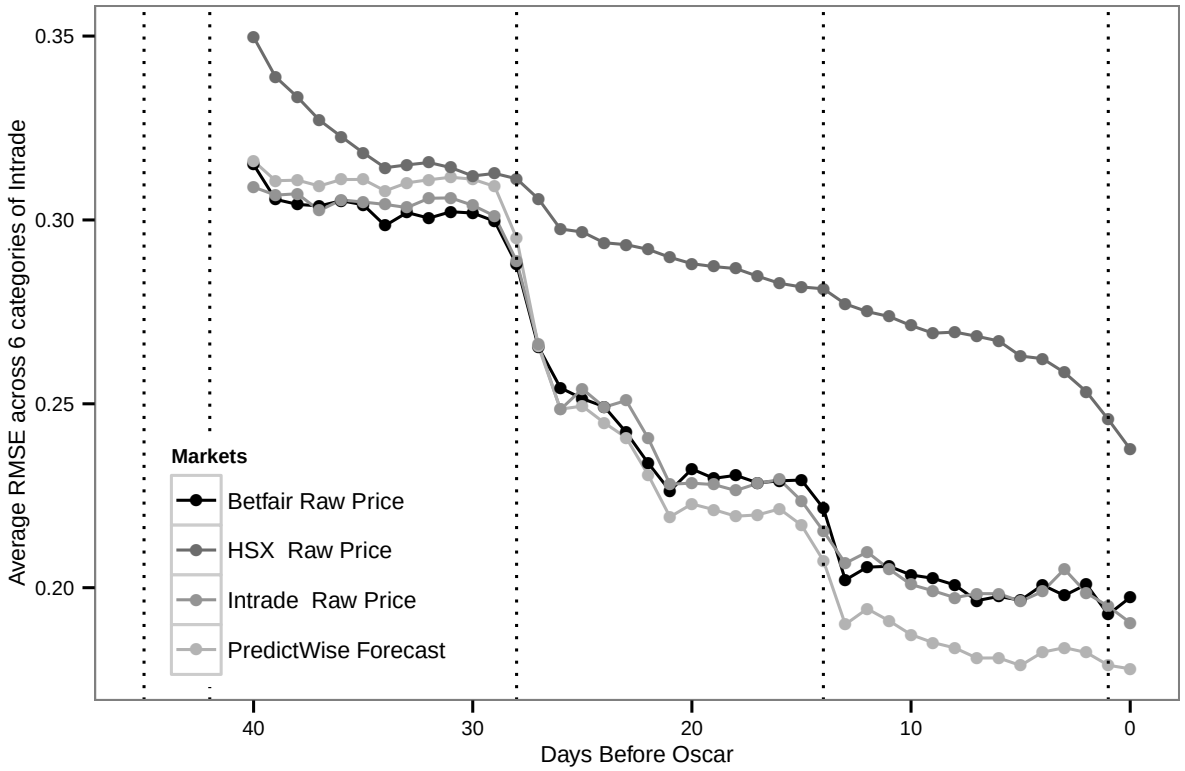

Figure 5: *RMSE of naive (raw price) forecast from prediction market and the PredictWise forecast data for all 6 common categories.* Error across all six common categories of Betfair, Intrade, HSX, and PredictWise for the 2013 Oscars. The awards shows are represented by vertical dotted lines, from left to right: Critic's Choice, Golden Globes, Guild, BAFTA, and Spirits.

We use the same methodology as with the polling and convert raw prices into translated prices by fitting a linearly varying coefficient $\beta(t) = \beta_0 + \beta_1 t$. Similarly to polling we reduced the number of outcomes in each category to five. In addition, in Sound Editing category, which had two winners, we assumed that the actual winner was *Zero Dark Thirty*.[8]

[8]We chose this approach due to the implementation convenience. A more correct approach would be to

The fitted translation function for Betfair has the form:

$$\beta(t) = 1.5146 + 0.0191t$$

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 crossvalidated standard errors for the coefficients are ± 0.0151 (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 time to the event decreases the raw price is increasingly undervalued in predicting the winner. This is likely due to the favorite-longshot bias being more pronounced as more prices reach the extremes and transaction costs and risk-loving behavior prevent the prices from reaching the underlying subjective probabilities of the traders.

In order to judge accuracy of any measurable outcome, it is useful to examine both the error and calibration. Figure 6 shows the comparative analysis of error across Betfair raw prices, translated probabilities and PredictWise. The figure shows the benefit of modeling raw prices as both translated Betfair and PredictWise have lower errors than the raw prices. PredictWise' model did well in the beginning of the cycle, but the translated Betfair pulls ahead later in the cycle. While we do not show this figure, translated Intrade is very similar to translated Betfair in overlapping categories and translated HSX has a much higher error over overlapping categories.

The forecasts are strong at calibration as well, which we check by charting the percentage of predictions that occur across the bins of predicted probabilities. Figure 7 demonstrates how well calibrated these outcomes are since all the points are in the vicinity of the perfect calibration line (the red dashed line). If the forecast calls for 20% likelihood of an outcome, it happens about 20% of the time. Again we temper the results of the translation by noting that these are within-sample, in that we have no additional years for out-of-sample evaluation. But, the coefficients used in the PredictWise forecast were derived from data in other domains and published ex-ante, showing how we can utilize calibration from other domains to make out-of-sample suggestions for the translations that work well.

2.6. Experts

Experts produce the forecasts that many stakeholders see, so it is important to understand them. We break the experts in this domain up into two categories. First, there are Oscar/movie pundits who use their domain specific experience and critical skills to discuss the event related outcomes. Second there are the experts of this age of "big data", who use statistical models on the quantifiable data for forecasting purposes.

Before the results are out, numerous pundits publish their critical reviews and likely winners onto the web. Their decisions are likely based on personal hunches or some undefined

use the weighted log likelihood with the two winners corresponding to separate observations with weights of 0.5. Since this is only one of 24 categories, the effect of this choice is negligible. *Zero Dark Thirty* was the leading candidate 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.

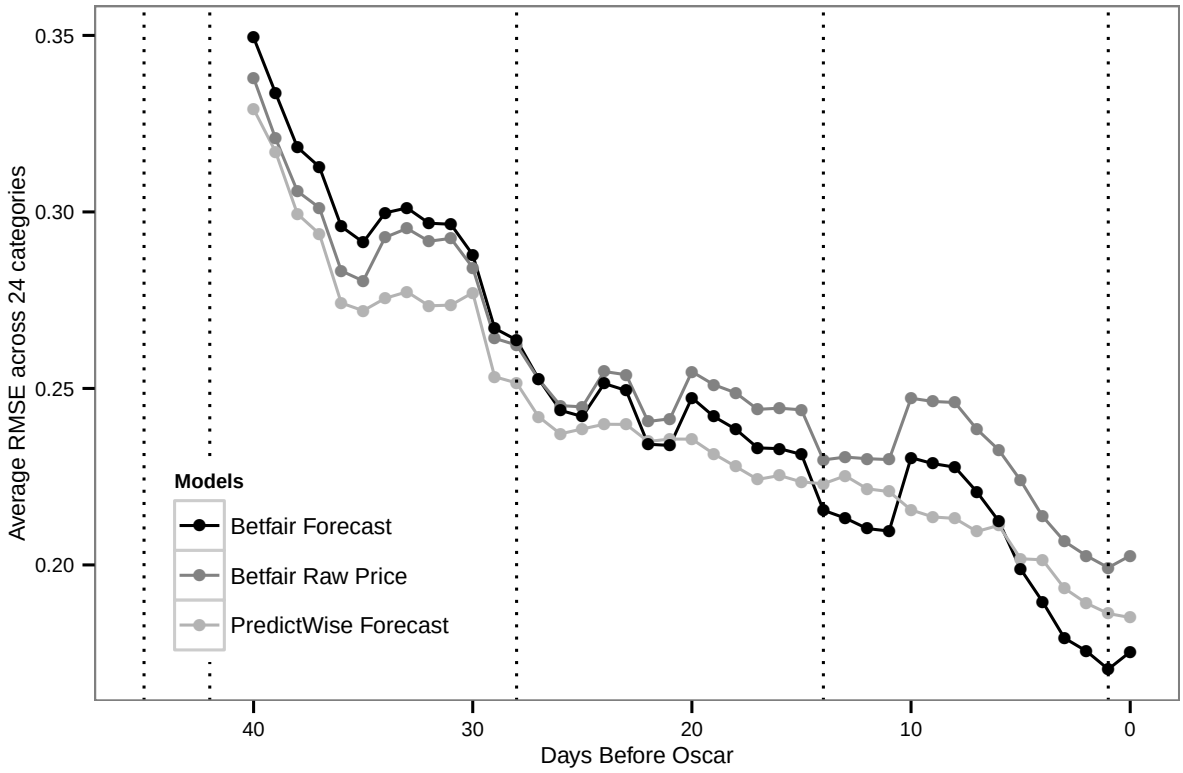

Figure 6: *RMSE of raw prices and translated prices from prediction markets for all 24 categories.* Error across all 24 categories using raw Betfair, translated Betfair, and PredictWise for the 2013 Oscars. The awards shows are represented by vertical dotted lines, from left to right: Critic's Choice, Golden Globes, Guild, BAFTA, and Spirits.

fundamental model. We referred to metacritic.com where they present aggregated (simple averages) probabilistic predictions from 40 different pundits and entertainment writers across all categories released on the web dated 21^{st} February 2013, 3 days before Oscar. A major drawback with this data source is timeliness as these are released few days before Oscar night.

The statistical experts make data based predictions and are basically similar to fundamental data sources. For example, Nate Silver of New York Times has presented his forecasts for three years 2009, 2011, 2013;[9] he publicly reveals the data and model, which is highly unusual for experts. He changes his model every year and this practice highlights two key concerns about expert forecast. First, there is a risk that late season changes in the data and methods can suffer from look-ahead bias (i.e., knowing the current expected outcome can invariably affect choices for data and methods leading a model to produce results that herd

[9]2009 (`http://nymag.com/movies/features/54335/`), 2011 (`http://carpetbagger.blogs.nytimes.com/2011/02/24/4-rules-to-win-your-oscar-pool`), and 2013 (`http://fivethirtyeight.blogs.nytimes.com/2013/02/22/oscar-predictions-election-style/`)

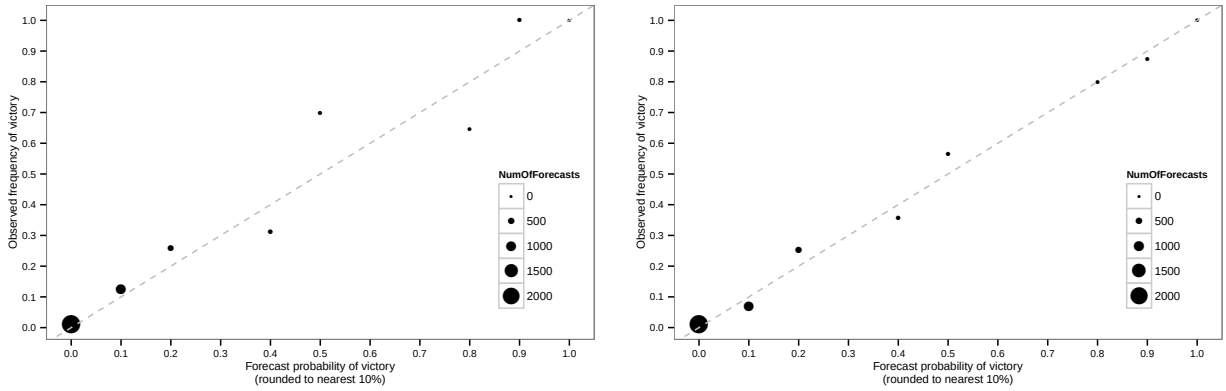

Figure 7: *Calibration of Betfair Raw Prices (left) versus Betfair Translated Prices (right).* Across all 24 categories using translated Betfair prices once per day for 40 days.

with others). Second, like most non-academics there is little protection from over-reliance on past results, which may not always work for future results (i.e., within-sample models that do not work out-of-sample). Initially in 2009, Silver used regression models over all relevant variables and 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 largely conforms to our findings on fundamental data. Beyond timeliness and accuracy, expert forecasts are not completely relevant as they rarely focus on a full catalog of categories. Ben Zauzmer is the next most cited expert we could find apart from Silver.

Table 4: *RMSE of expert forecasts.* Error across all categories forecasted by experts for the 2013 Oscars. Since experts are not constrained, after forecasting the six most prominent categories, we can assume that they forecast the easiest categories and do not forecast the hardest categories.

Experts	**Days before Oscar**	**Categories**	**Average RMSE**
Average of Oscar Pundits	3	24	0.2
Nate Silver	2	6	0.26
Ben Zauzmer	8-9	21	0.25

Table 4 presents the average root mean square error for 2013 forecasts by different experts Oscar pundits aggregate score, Nate Silver and Ben Zauzmer (fundamental data based) in order. It is crucial to note here that all three sources forecasts different categories, in generally the less categories, the easier the categories are on average. As the most likely categories to skip are the most obscure categories. Further, while the average of Oscar pundits does quite well, the average of the Oscar pundits is much better than the average Oscar pundit.

2.7. Comparison of All Methods

Fundamentals, prediction markets, and experts all have examples of forecasts in all 24 categories for 2013. Figure 8 compares their average RMSE over all 24 categories. Prediction

markets are significantly more accurate. 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 less corresponding awards and their average error right before the Oscars is 0.40, compared with the average error of 0.25 in 15 categories excluding those categories.

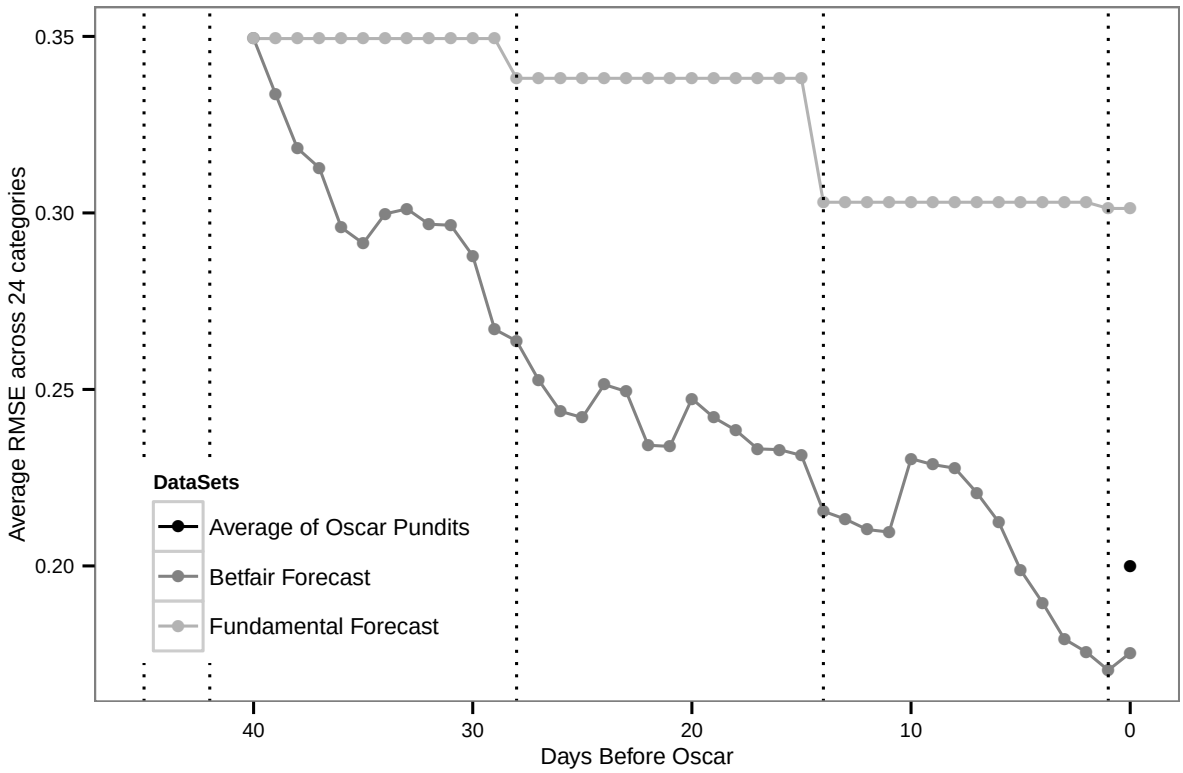

Figure 8: *Average RMSE of Oscar Pundits, translated Betfair prices and fundamental model for all 24 categories.* Error across all 24 categories using translated Betfair prices and the fundamental model for the 2013 Oscars. The awards shows are represented by vertical dotted lines, from left to right: Critic's Choice, Golden Globes, Guild, BAFTA, and Spirits.

Figure 9 shows the errors for 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, like 20 days out when there was a big social media push for self-selected respondents or one day out when there was a huge push of the polls to random respondents.

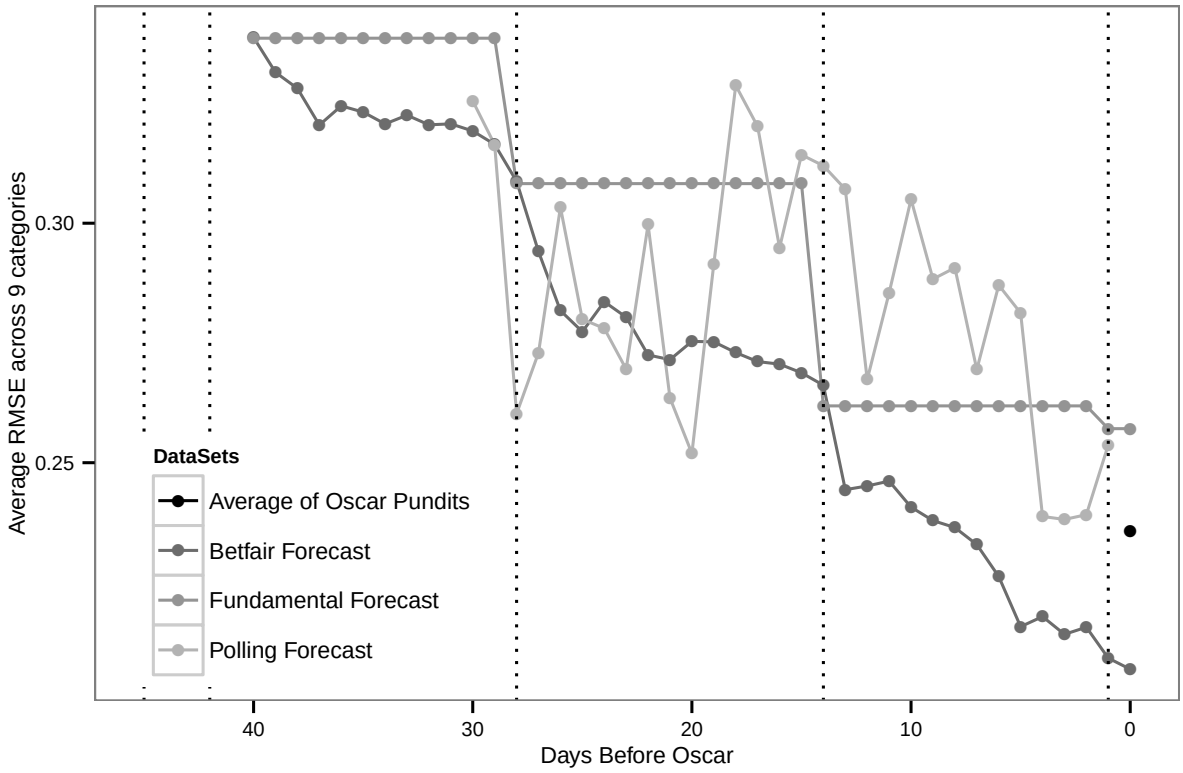

Figure 9: *Average RMSE of Oscar Pundits, translated Betfair prices, fundamental model, and translated polls for 9 common categories.* Error across all nine common categories using translated Betfair prices, translated polls (both types of users), and fundamental data for the 2013 Oscars. The awards shows are represented by vertical dotted lines, from left to right: Critic's Choice, Golden Globes, Guild, BAFTA, and Spirits.

3. Discussion

This paper provides new insights on the relative value of different forms of data for creating forecasts. The academic literature is prone to comparing forecasts on one dimension only, accuracy, but that is not adequate in any practical sense. First, stakeholders need to ensure the forecasts are relevant to their desired usage. For the Oscars that is probability of victory in all 24 categories for all nominees. Second, timeliness, meaning both early forecasts and frequent forecasts, are keys for efficient usage by stakeholders and researchers alike. For Oscars that means debuting the forecasts at the nominations and updating them continuously through the broadcast. Third, cost-effectiveness or the ability to scale the forecasts to new questions and domains is central to actual creation of the forecast. Finally, accuracy should not just mean a small error right before an event, but also robustness and calibration at other points in time.

Fundamental data is expensive to translate into a forecast and that cost is not commensurate with the timeliness and accuracy of the forecast. Fundamental data provides a

relevant forecast, as it can translate into forecasts for all 24 categories. But, in order to make all 24 forecasts there is an extraordinary cost in both data collection (which can be category specific) and modeling (which is specific to both categories and available data at any time). For timeliness the forecast needs to wait until the awards shows occur, so it is lacking most of its information at the nomination day. For 2013, this data is mostly collected by 13 days before the Oscars, after the BAFTA awards. Thus, unless it is judged by accuracy at its most opportune moment, it is relatively not accurate.

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 provide increased noise as the categories become more obscure, while the self-selected respondents may be informed in some of the obscure categories. The expectation question used in the polling is similar to implicit question in prediction markets, thus, it is not surprising the question works when it has informed users. But, polls do not provide the possibility of monetary reward that keeps respondents engaged on a continuous basis.

Prediction market based forecasts excel on all aspects: relevancy, timeliness, accuracy 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 they could be very accurate results, even if the liquidity is low for the more obscure categories. Prediction markets move in real-time as events unfold. And, they are extremely accurate: low errors and impressive calibration.

Expert forecasts are not as timely or accurate as other options.

This work should extend beyond the domain of the Oscars. All of the critiques we have of the varying data sources and their transformation into models confirm and expand on the body of literature noted in the introduction, which spans numerous domains.

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-NEW YORK THEN WASHINGTON- 320 (5878), 877.

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.

Guedj, O., Bouchaud, J.-P., 2005. Experts' earning forecasts: Bias, herding and gossamer information. International Journal of Theoretical and Applied Finance 8 (07), 933–946.

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., 2013. Fundamental models for forecasting elections. ResearchDMR. com/HummelRothschild_FundamentalModel.

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., 2013. Combining forecasts: Accurate, relevant, and timely. Available at http://www.researchdmr.com/RothschildForecast12.

Rothschild, D., Wolfers, J., 2011. Forecasting elections: Voter intentions versus expectations. Available at SSRN 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., 2013. Forecasting elections with non-representative polls. Available at http://5harad.com/papers/forecasting-with-nonrepresentative-polls.pdf.

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.

Appendix A. Coefficients in the Final Fundamental Model across All Categories

Table A.5: *Coefficients in the final fundamental model.* Standard errors provided in parentheses.

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 A.6: *Coefficients in the final fundamental model.* Standard errors provided in parentheses.

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)	0.1 (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 A.7: *Coefficients in the final fundamental model.* Standard errors provided in parentheses.

Variables	Cinematography	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)	-2.32 (1.131)	-2.625 (0.560)	-1.737 (1.135)	-1.378 (0.634)

Table A.8: *Coefficients in the final fundamental model.* Standard errors provided in parentheses.

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 (0.070)	——
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)