Non-Parametric Analysis of Social Influence Bias in the California Report Card

ABSTRACT
TODO

1. INTRODUCTION

TODO Social Influence is too broad replace with something more specific throughout

Online tools increasingly incorporate the widsom of crowds through participant ratings, opinions, and social media sharing. The responses from these tools, like traditional surveys, are subject to a variety of biasing tendencies which have been well studied[?]. A common feature of many of online tools is that they reveal aggregate statistics (eg. show the average rating for a product) before a participant shares their opinion. This allows the tool to be informative while collecting data. Crowdsourcing tools, in particular, have recently been applied in participatory democracy [?], where these aggregate statistics are often cited as an advantage over traditional opinion polling [?] as they increase the transparency of the system.

In recent work, Munchnik et al. [?], used a randomized experiment to determine the magnitude of *Social Influence Bias* in up-voting in Reddit.com. They randomly treated forum posts with extra up-votes and down-votes and measured the treatment effect; concluding that a statistically significant bias exists. The called the tendency *social herding* since subsequent users were statistically significantly more likely to agree with the aggregate statistics they saw.

In this paper, we explore a related problem in the topic Social Influence Bias of whether participants will actually change their submissions upon revealing an aggregate statistic after they submit their input. As a case study, we use the California Report Card [?], where participants graded the state of California on six timely issues. After submitting a grade, participants were shown the *median* grade of all other participants. We recorded any changes to previous grades that happened after the median was visible.

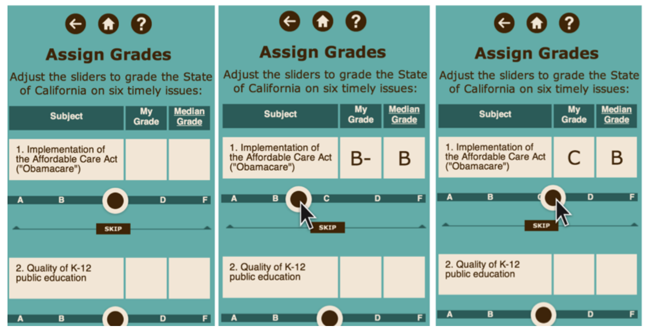

Figure 1: Grading in the California Report Card. Participants enter grades on six timely issues facing the State of California. After entering their grades, the median grade over all participants is revealed. Participants have the option to change their grades after seeing the median. We model the tendency to regress towards the medians.

The findings of Munchnik et al. would suggest that we will observe a biasing tendency in the form of a regression towards the observed median grade. They, however, explored this question only on a binary input mechanism (up or down vote) and we extend this analysis to grading sliders with 13 possible vales from (A+ to F). We further model this problem non-parametrically to make as few assumptions about the underlying distriubtion of grades. We use a variant of the Wilcoxon rank-sum test [?] to determine if there is a significant regression towards the median grades. In addition, this non-parametric analysis can be extended to compare the results with a randomized survey through SurveyMonkey and to consider the question of sequence-dependence [?].

In addition to the hypothesis testing, we fit models to the regression and its effects. We use an information theoretic criteria to fit a flexibile degree polynomial to model the regression. We also model how much more tightly centered around the median grade the changed grades are.

Our results suggest the following:

- There is a statistically significant regression towards the median for all of the issues.
- The grades are statistically significantly more concentrated around the median in comparision to a reference

survey.

- For 4 out of the 6 issues, this regression can be modeled as linear in the difference between the participant's initial grade and the median.

- For the other two issues, we found that the relationship can be modeled as a quadratic with greater regression for initial grades above the median.

2. RELATED WORK

Human opinion and rating in the presence of social influence has been well studied. In Asch's famous conformity experiments, participants were asked to match a line with a set of three three disparate lines, after 0 to 16 confederates had first given a unanimous incorrect answer. On average, 25% of participants conformed to the incorrect consensus. In 2011, Lorenz et. al described how social influence can undermine the effectiveness of crowd intelligence in estimation tasks. In particular, they describe social influence effect as causing a diminishment in diversity of opinion without improvement of its collective error.

Extending the study of social influence to online recommendation systems, Danescu-Niculescu-Mizil et. al's case study of Amazon helpfulness reviews argues that a written review's perceived helpfulness depends not just on the content of the review, but also the its score's relation to other scores. In order to better distinguish social influence from uninfluenced agreement, Muchnik et. al designed a randomized experiment in which comments were randomly up-treated or down-treated. Their work concluded a asymmetric herding effect, with a tendency towards positive bias. Sipos et. al argue that context along with an aggregate rating plays a large role in the users' ratings. That is, users may attempt to âĂIJcorrect" the average, by voting in a more polarizing manner (more positively or negatively).

The CRC presents a particular form of social influence in that the median grade is revealed after the user has adjusted the slider, and then given an opportunity to change. Zhu et. al examined a similar phenomenon in their experiment in which users rate their preference of an image, which was followed (either immediately or later) by a presentation of the crowd consensus opinion. Users were given an opportunity to change their response. Their work concluded the influence of conformity pressures is strongest, among other factors, when users are required to make their second decision sometime later, rather than immediately afterwards.

3. THE CALIFORNIA REPORT CARD

3.1 System Description

The California Report Card (CRC) is a web application that allows participants to advise the state government on timely policy issues. The CRC is divided into two phases: assessment and deliberation. In the assessment phase, participants grade the state's policies on a scale from A+ to F on six issues with a slider. Participants have the option to skip any issue. After a participant's first response, the median grade for all participants is revealed. The slider, is however, still active and participants have the option to change their grades. This process is illustrated in Figure 1. In the deliberation phase, participants submit textual suggestions on future issues to include in the report card. In this work, we focus on the assesment phase and defer an analysis of biases in the deliberation phase to future work.

3.1.1 The Six Issues

The six issues were:

- Implementation of the Affordable Care Act ("Obamacare")
- Quality of K-12 public education
- Affordability of state colleges and universities
- Access to state services for undocumented immigrants
- Laws and regulations regarding recreational marijuana
- Marriage rights for same-sex partners

These issues were posed in a constant sequential order with the same input scale (A+ to F). The issues were chosen to be timely and relevant to a majority of Californians.

3.2 Dataset and Experimental Setup

For each of the six issues, the report card collected around 1700 distinct inputs (both grades and skips). Grades were recorded every time the slider was released. The slider for the grades was discretized into 13 parts (A+, A, A-,...).

For analysis, we mapped these 13 grades onto a scale from 0 to 1, with 1 being an A+ and 0 being an F. For each participant p_j, we associate a 3-tuple of grades ($g_i[j]$, $m[j]$, $g_f[j]$) which reprsent the initial grade, median observed by the participant, and the final grade. To control for random changes or artifacts of the input device (eg. the slider stops), we counted changes that spanned a minimum time threshold of 3 seconds. With this definition, between 10% and 20% of the final grades involved at least one grade change. A detailed breakdown for each issue is illustrated in Figure 2.

We further conducted a reference survey through Survey Monkey asking the same questions (including the option to skip) without the median grade feedback. The reference survey had a sample size of 611 participants.

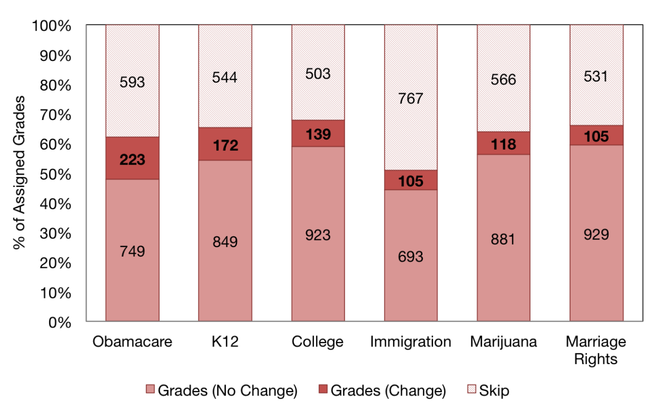

Figure 2: Breakdown of Activity in CRC Assesment.

4. HYPOTHESIS TESTING

Recall from the previous section that we define a 3-tuple for each participant: ($g_i[j]$, $m[j]$, $g_f[j]$). To test the biasing effect of the observed median m_i, our parameter of interest is the pearson correlation coefficient of the observed difference to the ultimate grade change: $\rho = corr(m_i - g_i, g_f - g_i)$. Testing this parameter of interest poses a few statistical challenges: (1) the discretization of the data leads to a multimodal distribution which are known to cause parametric statistical significance tests to perform poorly [?], (2) significant regression towards the median can be observed even if there is no biasing tendency, and (3) m_i changes over time.

To make challenge (2) more clear consider the following participant behavioral model. Suppose that participants are not acustomed to a slider-based input. We can model the first grade that the particpant leaves as uniformly randomly anywhere on the slider. As the participant begins to understand how to use the slider their use becomes more accurate, ultimately settling on a grade from our observed distribution of final grades. This model, the first grade is uniformly random and the second grade is a sample from the observed distribution, would result in a strong regression towards the median; even if there is no causal link.

Therefore, we avoid directly testing the correlation due to challenge (2), and propose an alternate parameter: the absolute deviations of the grades around the median. We propose a non-parametric model based on the Wilcoxon statistic [?] to test the hypothesis that the group of participants that changed their grades are more tightly centered around the median grade. To pass this test with significance, it is not enough that there is a regression towards the median from the initial grades, but also that the final grades are more concentrated than grades from those that did not change.

4.1 Non-parametric Significance Test

The test that we propose is related prior non-parametric and parametric tests such as the Seigel-Tukey test[?] and the F-Test [?] that test the spread of a distribution around a point such as the mean or the median. However, in our case, the median that participants observe changes over time. As the system collects more grades, it incrementally updates its median value.

Let P_n be the set of participants that did not change their grades and P_c be the set of participants that changed their grades. We define a set X_c, X_n of absolute deviations from the observed median of the final grade for each group:

$$X_c = \{|m[j] - g_f[j]|\} \ \forall j \in P_c \tag{1}$$

$$X_n = \{|m[j] - g_f[j]|\} \ \forall j \in P_n \tag{2}$$

Now, for the set X_c, we calculate the Wilcoxon rank-sum statistic. We assign a rank to each of the absolute deviations in the union set $\mathbf{X} = X_c \cup X_n$ (ie. the largest change has rank 1 and the smallest has rank $|X_c \cup X_n|$. For X_c, we sum the ranks of the deviations within its set:

$$W_c = \sum_{j \in P_c} R_j \tag{3}$$

Under the null hypothesis $median(X_n) = median(X_c)$, the ranks will be evenly distributed between each group. Therefore, the null expected value and variance of W is:

$$\mathbb{E}(W) = \frac{(|\mathbf{X}| + 1) \cdot |X_c|}{2} \tag{4}$$

$$var(W) = \frac{(|\mathbf{X}| + 1) \cdot |X_c| \cdot |X_n|}{12} \tag{5}$$

For the significance level α, we can test the probability that our calcuated W_c comes from the null distribution. A significant result means that for the participants that changed their grades the changed changes are more tightly centered around the median grade they observed.

The same analysis can be extended to test X_c against the initial absolute deviations for the change group X'_c:

$$X'_c = \{|m[j] - g_i[j]|\} \ \forall j \in P_c \tag{6}$$

4.2 Justification For The Wilcoxon Statistic

In Figure 3, we show the distribution of absolute deviations for the Marriage Rights issue. We see that the distribution

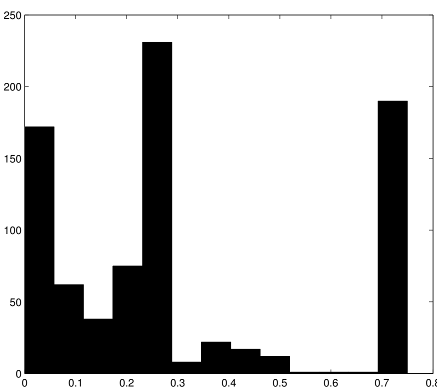

Figure 3: TODO

is multimodal and discrete. Parametric tests such as the z-test and the t-test have been shown to have weaker statistical power in many families of multimodal distributions such as mixtures of Gaussians [?]. Rank-based tests tend to be more robust to the multimodality and in fact don't depend on the actual values on the relative frequency of the ranks in the test set.

5. SEQUENCE DEPENDENCE

In the CRC, we pose each of the six issues in a constant order. In this section, we develop a model for testing the effect of the sequence of ratings. In [?], the authors suggest that *path depedence* is significant in online tools. In particular, we test to see whether deviation from the median on previous issues is correlated with deviation on the current issue. We test the following hypothesis: if participants observe that their grades significantly deviate from the median on previous questions, their future responses will be more tightly centered around the median.

This hypothesis is challenging to test as responses to issues may be correlated; even excluding the bias. Consider the following example, if the grades are positively correlated, then low grades on one question could imply even lower grades on another. In this case, we would see an increase in deviations even though it is not attributable to the biasing tendency. Consequently, we build a model that compares the CRC to the SurveyMonkey reference survey. We test to see if the relationship between the deviation of a participant's past grades and their current grades is different between the CRC and reference survey.

Let d_{ij} be the absolute deviation from the median grade of participant j's grade on issue i. We define a statistic P_{ij}, which is the mean of all of the absolute deviations on the previous issues:

$$P_{ij} = \frac{1}{i-1} \sum_{k<i} d_{kj} \tag{7}$$

For each issue $i > 1$, we can get a set of differences between the absolute deviation of the current issue and the average previous absolute deviations:

$$D = \{(P_{ij} - d_{ij})\} \forall j \tag{8}$$

We can calculate the same statistic D_r for deviations for the reference survey. For a given issue, these two sets illustrate the trend in deviations from the median. A large positive value implies that a participant who disagreed greatly with the median grade before is now much closer to the median. Conversely, a negative value implies their response deviates more.

While this statistic is difficult to intepret for an individual participant as their assesments may vary issue to issue, we can compare the distributions of differences from the CRC and Reference Survey. The two sets can be tested with the Wilcoxon model in the previous section. The results in [?] suggest that the CRC should show larger differences; corresponding to increasingly moderate grades by participants who observed that they disagreed with the median in the past. Thus, we test to see if the differences in the set from the CRC D are statistically significantly higher than those from the reference survey D_r. The Wilcoxon testing procedure is the following: (1) we rank the differences in $D \cup D_r$, (2) we calculate W which is the sum of the ranks in D, and (3) using the equation from the previous section we test the calculated W under the null hypothesis distribution.

A significant result means that in comparision to the reference survey, CRC participants future responses were more concentrated around the median (ie. a higher difference between $P_{ij} - d_{ij}$. This test is particularly interesting in the context of initial grades rather than final ones. We can test to see how the concentration of grades around the median changes even without the biasing effect of revealing the median, and whether participants have a tendency to *guess* the median grade.

6. PARAMETER ESTIMATION FOR GRADE CHANGE MODEL

In the previous sections, we proposed a technique to test the significance of the regression towards the median. In this section, we build a model to describe the relationship between the variables in the 3-tuple ($g_i[j]$, $m[j]$, $g_f[j]$). We also contrasted two different parameters of interest: correlation and absolute deviation. In this section, we will further build on this to estimate two quantities: a functional relationship between $m[j] - g_i[j]$ and $g_f[j] - g_i[j]$, and a quantification of how much more concentrated the changed grades are Δ. The functional relationship, related to the correlation, will tell us how to predict a final grade given an observed median. The Δ parameter will tell us how much more tightly grouped around the median the final grades are.

6.1 Modeling Heterogenous Changes

Previous work, suggests that Social Influence bias is not homogenous; that is a negative influence is different in magnitude than a positive influence [?]. This means that we cannot assume that the relationship between $m[j] - g_i[j]$ and $g_f[j] - g_i[j]$ is linear.

Similar to the previous section where we applied non-parametric tests, we propose a information theoretic model search that allows flexible parameter selection without making strong assumptions about the nature of the relationship. Let $f \in \mathcal{P}^k$ be a polynomial of degree k. The square loss of f, is the error in predicting $g_f[j] - g_i[j]$ from $f(m[j] - g_i[j])$:

$$\mathcal{L}(X_c; f, k) = \sum_j ((g_f[j] - g_i[j]) - f(m[j] - g_i[j]))^2 \tag{9}$$

For a given k, the best-fit polynomial minimizes this square-loss:

$$f_k^* = \arg\min_f \mathcal{L}(X_c; f, k) \tag{10}$$

To search over the space of polynomial models, we apply a well-studied technique called the Bayesian Information Criterion (BIC) [?]. This penalty can be interpreted as bias towards lower degree models, in other words, an Occam's Razor prior belief. This we reformulate the optimization problem in the following way to incorporate the BIC:

$$\arg\min_{f,k} |X_c| \log(\mathcal{L}(X_c; f, k)) + k \log(|X_c|) \tag{11}$$

The resulting optimal polynomal will tell how the regression affects varies as a function of $m[j] - g_i[j]$ while controlling for over-fitting to our data. This optimization problem is non-convex so we incrementally try polynomials of degree 1,2,3.. etc. until we reach a local minimum.

6.2 Concentration of Grades

We can further estimate how much more concentrated changed grades are around the observed median. Recall in Section 4, we tested the significance of the absolute deviations using a Wilcoxon test statistic. The Wilcoxon statistic can be inverted to estimate a most likely *shift parameter*, that a constant shift Δ in the distribution of absolute deviations X_c that maximally aligns them with X_n (ie. $X_c + \Delta$ is most supported by the null hypothesis). Since X_c is a set of absoulte deviations, Δ tells us how much more concentrated X_c is than X_n around the observed medians. This parameter is relevant to the design of recommendation algorithms use proximity (eg. clustering or nearest neighbors).

We refer to [?] on the derivation of Δ and its confidence interval:

$$D = \{x_n[j] - x_c[i]\} \ \forall i, j \in X_n, X_c \tag{12}$$

$$\Delta = median(D) \quad (13)$$

7. RESULTS

7.1 Observed Regression Towards the Median

In Figure 4, we plot $m_i - g_i$ (the observed difference) against $g_f - g_i$ (change in grade) for those participants that changed their grades. Supporting our initial hypothesis, we find that the values are positively correlated, which we define as a change towards the median.

Issue	N	Corr	P-Value
Obamacare	223	0.4580	5.1270e-20
K12	172	0.4813	1.6573e-18
College	139	0.4263	3.9772e-13
Immigration	105	0.5856	1.8984e-20
Marijuana	118	0.5397	3.0882e-19
Marriage Rights	105	0.5538	4.0921e-25

Furthermore, the correlations are significant with respect to the null correlation hypothesis. The significance test shows it is highly unlikely that there is no correlation between the observed difference and the change in grade. Note that we discussed that this correlation does not on its own imply a tendency to regress towards the median grade as disccused in Section 4, as other models could result in similar correlations.

7.2 Non-Parametric Test Of Distance From the Median

We applied the non-parametric test proposed in Section 4. Figure 9 shows the mean absolute deviation for each group, and Table 7.2 tests its significance.

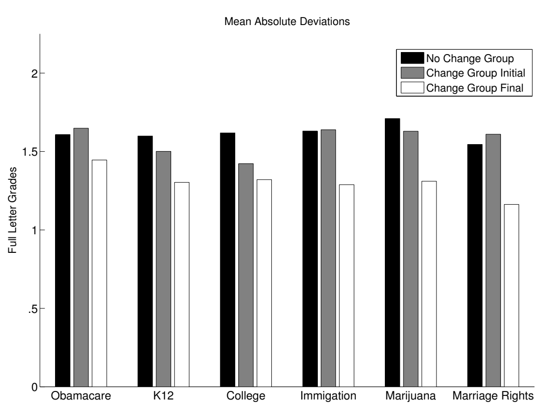

Figure 5: TODO

Issue	P (X_c vs. X_n)	P (X'_c vs. X_c)
Obamacare	0.0286	0.0161
K12	2.1314e-06	0.0086
College	1.3033e-04	0.0415
Immigration	7.3456e-07	4.4170e-05
Marijuana	2.7549e-10	4.2560e-05
Marriage Rights	3.5946e-06	2.4644e-10

For all of the issues, we find that set of absolute deviations from the median X_c is statistically significantly smaller compared to both X_n and X'_c. This suggests that the participants that changed their grades tended to be more tightly centered around median grade.

7.3 Sequence Dependence

Using the model proposed in Section 5, we calculated the test statistics for both the CRC and the Reference Survey. We found that for all issues the statistic was higher for the CRC suggesting an effect corroborating results in other work such as [?]. However, none of the results passed a $p < 0.05$ statistical significance test. We believe that these results

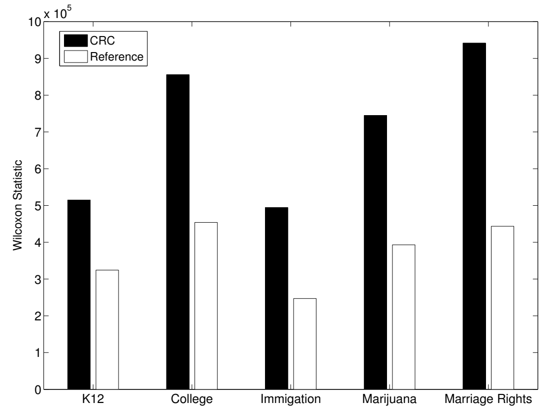

Figure 6: TODO

suggest that there is some sequence dependence in the CRC, however, we cannot definitively conclude that from the current quantity of data.

7.4 Grade Change Model

In Figure 7 and Figure 8, we show the results of our model search and locally optimal model for each issue. We found for four out of the six issues, K12, College, Immigration, and Marijuana, the model we found was linear. However, for Obamacare and Marriage Rights, we found that the relationship was quadratic.

Figure 8 illustrates the nature of the quadratic relationship, and we see heterogeneity between a positive regression towards the median and a negative regression. Participants who initially graded the state higher than the median had a more significant tendency to regress downwards. This result is interesting for a few reasons: (1) contrary to our initial expectations the relationship is largely linear and (2) non-linearities appear in the two issues that received the highest grades which also happen to be highly politicized issues. There are many possible explanations for this including non-response bias [?] or aversive response [?]; and we defer a more detailed analysis to future work.

7.5 Shift-parameter estimation

In Figure 9, we show the results for the Δ parameter estimate from inverting our hypothesis test. We find that on average grades in our change group were about half a letter

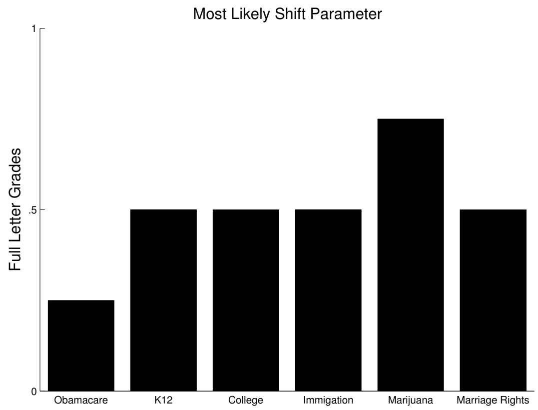

Figure 9: TODO

grade (ie. +, -) closer to the median than grades from participants who didn't change. This parameter is relevant to both recommender systems and prediction tools. In the simplest case, if we were to build a grade prediction tool that simply predicted the median grade, we could get misleadingly low prediction error. Likewise, algorithms that rely on proximity such as clustering or k-nearest neighbors could be misled to create a single big cluster around the median, when in fact the clustering around the median may be due to a biasing tendency.

7.6 Comparision To Reference Survey

We applied our proposed non-parametric test to compare the absolute deviations in the group of participants who changed their grades in the CRC with results from a reference survey. For the reference survey, we calculated the absolute deviation around the median (which the participants were not shown). We found for all but one issue the grades from the CRC were statistically significantly closer to the median than ones from the reference survey.

Issue	Med(Ref)	Med(CRC)	p-val
Obamacare	B	B	0.0078
K12	C+	C	0.3563
College	C-	C-	0.0011
Immigration	C	C+	0.0277
Marijuana	C	C	0.0076
Marriage Rights	B+	B+	0.0494

Furthermore, the two surveys aligned nearly perfectly in aggregate.

8. FUTURE WORK

The methods we proposed have several interesting directions of future interest. We want to extend our work to quantify biases in textual data. The California Report Card collects textual suggestions from participants in addition to the quantiative assesment results. Participants are encouraged to read the responses of others before leaving a suggestion of their own. We suspect that this may lead to a bias in the topics discussed by participants, and we would like to explore how similar non-parametric models can be extended to textual data.

Another compelling direction is to attempt to parameterize our model. We will explore whether we can model the grades as a mixture of binomial distributions (a discrete analog of a mixture of gaussians), and try to derive optimal tests and models for this data. Intuitively, parametrization should lead to increased statistical power and better fitting models; assuming that the data fits the underlying parametrization.

9. CONCLUSION

We proposed non-parametric hypothesis tests and models to evaluate the biasing tendency of visible aggregate statistics in the California Report Card. We found that revealing the median led to a statistically significantly tighter grouping of grades around the shown median grade.

We modeled the biasing effect as a regression towards the median grade and fit polynomial to represent the functional relationship between a participant's observed difference with the median and then subsequent grade change. We applied an information theoretic criteria to select a model of appropriate complexity. We found that this relationship was quadratic in two out of the six issues, representing a heterogeneity in biasing for positive and negative differences with the median. We further showed how non-parametric ideas could be extended to the problem of Wilcoxon shift parameter estimation and quantify the effects of the biasing tendency.

In principle, the methods we proposed can be applied to test and model biases in a wide variety input mechanisms. This is a key motivation for our non-parametric approach. Understanding these biases, can give insight into the behavior of recommender systems that train on such data.

10. ACKNOWLEDGMENTS

APPENDIX

A. HEADINGS IN APPENDICES

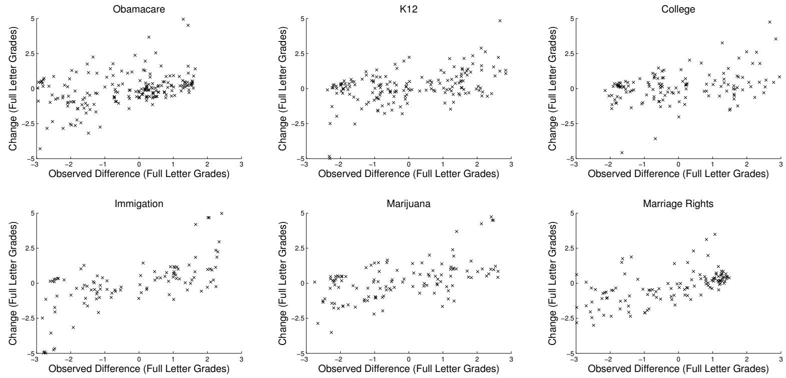

Figure 4: TODO

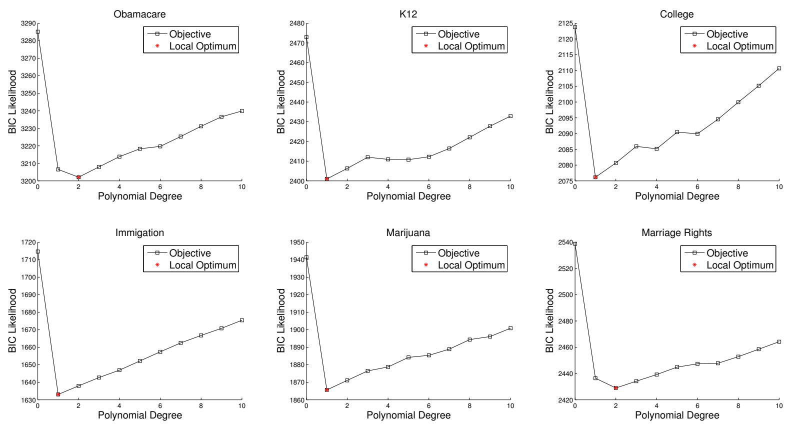

Figure 7: TODO

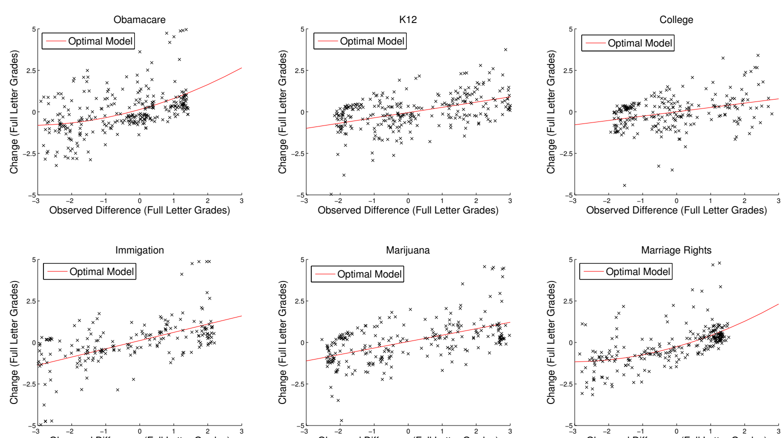

Figure 8: TODO