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

ABSTRACT

1. INTRODUCTION

User-assigned ratings are a key component of almost all recommender systems. Rating data, like traditional surveys, are subject to a variety of biasing tendencies [11]. We explore a class of biases, collectively called *Social Influence Bias*, which arise from feedback from the actions of other users in the system [9, 15, 21]. One aspect of Social Influence Bias is the phenomenon called *Social Herding*, where the feedback from the community encourages future participants to conform to what they perceive as the "norm" in the community. In a rating system, this can lead to an increased tendency to leave ratings close to the mean or the median rating. The effects of social herding are crucial to the design of recommendation algorithms as many algorithms assume statistical independence between different users and use the spatial relationships between numerical features representing those users.

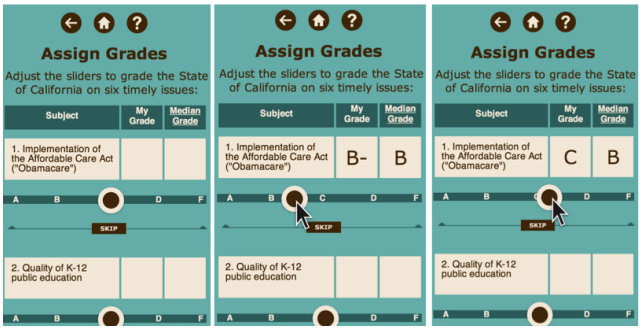

Figure 1: Grading in the California Report Card. Participants enter grades on six timely issues facing the State of California. After entering their grades, the median grade over all participants is revealed. Participants have the option to change their grades after seeing the median. We model the tendency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

A common feedback mechanism is the use of aggregate statistics, for example, showing the average rating for a product before a participant shares his or her rating (Figure 1). In many cases, such as product reviews, it is not practical to hide this information from potential raters. The use of social content is an established user experience design technique to incentivize participation and increase user engagement with the tool [19]. Furthermore, an application of particular interest is online participatory democracy where open aggregate results increase the transparency of the system [1, 18, 17].

In recent related work, Muchnik et al. [16], used a randomized experiment to determine the magnitude of social herding 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 social herding tendency exists. We study a related effect using a new platform, the California Report Card (CRC), which reveals the median values to users *after* they assign their own ratings, and then allows them to modify their ratings. We test whether users who have already submitted a rating will actually *change* their existing ratings upon learning the median rating for the population and if there is a tendency to "herd" towards the median grade.

For comparision with the CRC, we ran a reference survey through SurveyMonkey which was given to a random sample (N=611) from the company's paid pool of California participants. In this survey, we asked users to respond to the same questions as the CRC on the same grading scale; but without the feedback of the median grade.

The findings of Muchnik et al. suggest that social herding will be observed in the form of a regression towards the observed median grade during the grade changes. In this paper, we test the hypothesis of social herding, propose a model for the relationship between observed medians and subsequent grade change, and provide experimental results comparing the data from the CRC to the randomized reference survey.

1.1 Hypotheses and Contributions

Null Hypothesis Viewing the median grade does not affect how a user chooses to change his or her grade, and does not affect any future grades given by the participant.

Social Herding Towards the Median Grade changes are

on average towards the median grade. Also, the final grades of participants who change their grades are more tightly concentrated around the median from participants who did not change their grades and participants from the reference survey.

Social Herding and Question Order Disagreement with the median on previous questions affect how a participant grades future questions. Participants who disagreed with the median greatly are more likely to leave future responses that are closer to the median.

We develop non-parametric testing procedure, based on the Wilcoxon Rank-Sum statistic (also called the Mann-Whitney statistic) [13], to test these hypotheses. We chose a non-parametric framework because Muchnik et al. focused on only a binary input mechanism (up or down vote) and we extended this analysis to grading sliders with 13 possible vales from (A+ to F) without having to make strong assumptions about the distribution of grades. In addition to the hypothesis testing, we model grade changes with a polynomial regression. As before, to avoid having to make a strong assumption about the structure of the model, we use an information theoretic model to learn a flexibile degree polynomial.

2. RELATED WORK

In Asch's famous conformity experiments [3, 2, 6], groups of participants were asked to match a line with a set of three different sized lines one of which was of the correct size. In reality, only one of the participants was "real" and the others were actors who unanimously chose an incorrect choice. On average, 25% of participants conformed to the incorrect consensus compared to 1% of incorrect answers in a control group.

The Asch model for conformity is the theoretical basis for social herding [4, 5], and herding has been a popular consumer choice model in economics [7, 10, 12]. Such models have also been studied in psychology as "persuasion bias" [9]. In 2011, Lorenz et al. described how these biases can undermine the effectiveness of crowd intelligence in estimation tasks [14]. They argue that social herding causes a diminished diversity of opinion potentially leading to inefficiencies and inaccurate collective estimates. Danescu-Niculescu-Mizil et al. analyze helpfulness ratings on Amazon product reviews [8]. They found that the helpfulness ratings did not just depend on the content of the review but also its aggregate score and its relationship to other scores. In order to better distinguish social influence from other biases, Muchnik et. al designed a randomized experiment in which comments on Reddit.com were randomly up-treated or down-treated [16]. They concluded a statistically significant bias where a positive treatment increased the likelihood of positive ratings by 32%. In both Danescu-Niculescu-Mizil et al. and Muchnik et al., they looked at the problem of Social Influence bias in an a priori setting, where users see the aggregate statistic before giving their rating. Our work tests for a particular form of social influence where users are given the opportunity to change their opinions following the feedback.

Zhu et al. conducted an experiment in which users evaluate an image on a subjective question with binary scale (eg. "Is this image cute?"), which was followed (either immediately or later) by a presentation of the crowd consensus opinion [22]. Users were given an opportunity to change their response, and they concluded that there was a significant tendency to change submissions. The tendency to change was the strongest when users were asked to make their second decision much later and not immediately after the first. However, Zhu et al. also acknowlege there are competing psychological factors at work in this experiment. Along these lines, 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 "correct" the average, by voting in a more polarizing manner (more positively or negatively) [20]. We extend this prior work to measure and predict these changes when the input is more complex than a binary scale, and propose a non-parametric methodology that can be, in principle, extended to a variety of different input mechanisms. Our model can also account for a changing aggregate statistic such as a median rating changing as more data is collected.

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. When participants arrive at the application, they "grade" the state on following six issues:

- 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

Grades were assigned on a thirteen point scale (A+,A,A-,...,D-,F). These issues were posed in a fixed sequential order each with the same input scale. Particpants submitted grades using a click-and-drag slider interface as illustrated in Figure 1. On mobile devices this slider required the participants to touch and drag their finger to the desired grade.

Upon release of the slider, the CRC reveals the median grade for that issue over all prior participants. Even after the median grade is revealed the slider is still active and participants can change their grades. However, it is important to note that participants were not explicitly told that they could change their grades. Another important observation is that participants who accessed the application at different times may have seen different median grades as they were calculated based on the data upto that point. We recorded the initial grade, the median that the participant observed, and any subsequent changes along with timestamps for each of the events. Grading all of the six issues was not manditory and participants had the option to skip any of the issues.

The CRC has an additional open-ended discussion phase where participants submit textual suggestions on future issues to include in the report card. In this work, we focus on the first phase and defer an analysis of biases in the discussion phase to future work.

3.2 Notation

To analyze this data, we mapped these 13 grades onto a scale from 0 to 1, with 1 being an A+ and 0 being an F. Let P denote the set of all participants. For each participant $p_j \in P$, 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. For each issue, we divided the participants into three subsets of P: ones who did not change their grades P_n, ones who changed P_c, and ones who skipped the question P_s. Our primary objective is to test the distributional properties of rating tuples from participants in P_n compared to those in P_c.

To ensure that all participants in the set P_c had an opportunity to see the median grade and then react, we filtered this group using the timestamps. The median grade appears in the interface with an animation whose completion time varied between devices, so we set a grace period of 3 seconds before we categorized the participant into set P_c.

For consistency, we use the same notation to describe participants in the reference survey. We denote the set of reference survey participants as set R, and each participant is associated with a 3-tuple $(g_i[j], m[j], g_f[j])$. However, since the reference survey does not reveal the aggregate statistics $g_i[j] = g_f[j]$ and $m[j]$ is the median of the prior participants (which is not shown).

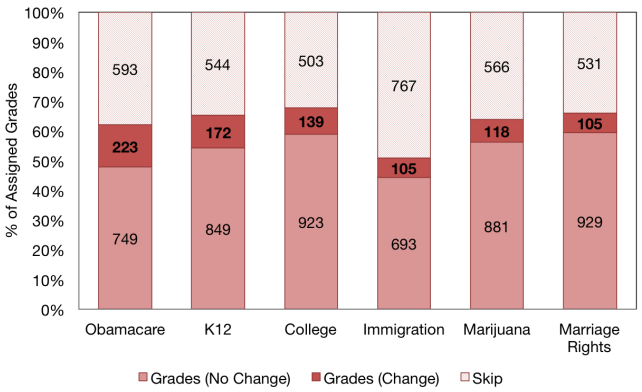

Figure 2: Breakdown of Activity in CRC Assesment.

4. SOCIAL HERDING TOWARDS THE MEDIAN HYPOTHESIS TEST

There are three principle challenges in testing **Social Herding Towards the Median** hypothesis. The first challenge is that parametric significance tests comparing two sample means such as the two sample t-test and z-test are known to perform poorly for multimodal distributions. Another significance test that is commonly applied to compare spreads of distributions is the F-test, which is also known to perform poorly for many non-normal distributions [**?**]. Furthermore, this test is usually used to test the spread of data around the mean, which only in very special conditions such as normal distributions, aligns with the median. The discretness of our data leads to mutli-modal distributions which are not optimal for these testing methods.

The second challenge is that there is a natural tendency for grades to concentrate around the median even without a bias. 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 participant 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 with seeing the median.

Finally, the median grade m_i can be different for each participant. The median grade is calculated over all prior participants and thus is dependent on when the participant submitted their first grade. In practice, the median will eventually converge for a large number of participants, but it would be incorrect to measure concentration around a final median.

To address these three challenges, 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 that the participants observed. Our tests compare absolute deviations around the median for P_n, P_c, and R; which, as a relative test, controls for the natural tendency for grades group around the median. Furthermore, it is more robust to the effects of alternate models such as the one described in our second challenge in comparision to a direct test of correlation.

4.1 Non-parametric Significance Test

Recall that P_n is 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:

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

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

For the purposes of hypothesis testing herding behavior, we ignore the sign of the deviation. However, in Section 6, where we build a predictive model for the changes, we include the sign.

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:

$$W_c = \sum_{j \in P_c} R_j \quad (3)$$

The **Null Hypothesis** is that absolute deviations in X_c are the same size as X_n. Under this 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. In other words, the test calculates the probability that a random subset of users (ignoring the categorization P_n and P_c) can have the observed difference in rank-sum values. 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 used to test X_c against the absolute deviations in the reference survey X_r

$$X_r = \{|m[j] - g_i[j]|\} \; \forall j \in R \tag{6}$$

or for initial vs. final ratings in the change group X_c':

$$X_c' = \{|m[j] - g_i[j]|\} \; \forall j \in P_c \tag{7}$$

4.2 Quantifying Concentration of Grades

In addition to testing the hypothesis, we can also quantify the effects of social herding. 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 [13] on the derivation of Δ and its confidence interval:

$$D = \{x_n[j] - x_c[i]\} \; \forall i, j \in X_n, X_c \tag{8}$$

$$\Delta = median(D) \tag{9}$$

5. SOCIAL HERDING AND QUESTION ORDER

In this section, we develop a model for testing the effect of the sequence of ratings. Order effects have been well studied in surveying [**?**], and we look at order effects in the context of social herding. Recall, that we posed the each of the six questions in a fixed order. Our question of interest is: given a participant's average disagreement with the median grade (measured by the absolute deviation) on the previous issues, how is their grade to the following issue affected? We hypothesize that participants will become more moderate in their grades if they observe that their grades a consistently in disagreement with the population consensus.

This hypothesis is challenging to test as responses to issues may be correlated; even excluding any form of 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_{kj} be the absolute deviation from the median grade of participant j's grade on issue k. We define a statistic $\bar{d}_{kj}$, which is the mean of all of the absolute deviations on the previous issues:

$$\bar{d}_{kj} = \frac{1}{k-1}\sum_{l<k} d_{lj} \tag{10}$$

If an issue was skipped by participant j, we exclude it from the average. For each issue $k > 1$, we can get a set of differences between the absolute deviation of the current issue and $\bar{d}_{kj}$:

$$D_k = \{(\bar{d}_{kj} - d_{kj})\} \forall j \tag{11}$$

We can calculate the same set of deviations for the reference survey which we call $D_k^{(r)}$. When the differences in D_k are on average positive it means that on issue k participants were more moderate than previous issues and vice versa if the differences are negative. So formally, our hypothesis test compares whether the set of differences in the CRC D_k are larger than the set of differences in the reference survey $D_k^{(r)}$. A significant result means that in comparision to the reference survey, CRC participants showed a greater tendency to center their grades around the median after disagreeing on previous issues.

We can apply the same Wilcoxon rank-sum model discussed in the previous section to test this hypothesis. The testing procedure is the following: (1) we rank the differences in $D_k \cup D_k^{(r)}$, (2) we calculate W which is the sum of the ranks in D_k, and (3) using the equation from the previous section we test the calculated W under the null hypothesis distribution. The null distribution models the null hypothesis that there is no difference between D_k and $D_k^{(r)}$, and given this hypothesis what is the probability we will observe the rank-sum statistic W.

This test is particularly interesting in the context of initial grades. If we construct our set D_k so that $\bar{d}_{kj}$ is based on final grades and d_{kj} is the deviation of the initial grade, we can test to see how the concentration of grades around the median changes even without the biasing effect of revealing the median. The implications of this question are interesting since this tests whether participants have a tendency to *guess* the median grade after prior disagreement with the median.

6. PREDICTING MAGNITUDE OF GRADE CHANGES

In the previous two sections, we proposed a technique to test the significance of the social herding hypothesis. In this section, we build a model to describe the relationship between the variables in the 3-tuple ($g_i[j]$, $m[j]$, $g_f[j]$). In other words, given a participant's current grade, the median they observed, can we predict the final grade.

6.1 Modeling Changes

Previous work, suggests that social herding is not a homogenous effect, namely, positive influences are different from negative influences. In Muchnik et al. [16], they found that when they positively treated posts with higher up-vote counts in Reddit it lead to a significant increase in the likelihood of additional up votes (32% more likely). On the other hand, they argue negative treatments inspired correction behavior; where some participants wanted to correct what they felt was an incorrect score. They found that this also increased the likelihood of up-voting (88% more likely).

These results suggest that the effects of social herding can be non-linear and are very context/question dependent. Similar to the previous section where we applied non-parametric tests that did not make a strong assumption about the distribution of the data, we propose a information theoretic model search that allows flexible parameter selection without making strong assumptions about the nature of the relationship. Conditioned on the event that the participant changes their grade, we learn a functional relationship between the observed median and initial grade that can be a polynomial of any degree. While the space of all polynomial models is fairly exhaustive, we acknowledge that this model can only fit curves that are continous and smooth.

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 $f(m[j] - g_i[j])$:

$$\mathcal{L}(X_c; f, k) = \sum_j ((g_f[j] - g_i[j]) - f(m[j] - g_i[j]))^2 \quad (12)$$

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

$$f_k^* = \arg\min_f \mathcal{L}(X_c; f, k) \quad (13)$$

For a given k, this problem can be solved with least squares. To search over the space of polynomial models, we apply a well-studied technique called the Bayesian Information Criterion (BIC) [**?**, **?**]. This technique converts the optimization problem into a penalized problem that jointly optimizes over the "complexity parameter" k. This penalty can be interpreted as bias towards lower degree models, in other words, an Occam's Razor prior belief. Cross-validation is an alternate method to empirically determine optimal model, and in practice, they give very similar results. BIC, however, is derived through maximum likelihood estimate and is not an empirical so the learned model has a notion of optimality conditioned on the BIC prior belief.

Thus, we reformulate the optimization problem in the following way to incorporate the BIC penalty:

$$\arg\min_{f,k} |X_c| \log(\mathcal{L}(X_c; f, k)) + k \log(|X_c|) \quad (14)$$

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. In general, this optimization problem is non-convex so we incrementally try polynomials of degree 1,2,3.. etc. until we reach a local minimum.

7. RESULTS

We evaluated our models on data collected from the California Report Card between January 18th to April 20th. We adminstered our reference survey through SurveyMonkey between March 8th and March 14th. We consider a set of 1575 total participants from the CRC and a sample of 611 SurveyMonkey participants whose grading activity was as follows:

Issue	No Change	Change	Skip	Median
CRC				
Obamacare	749	223	593	B
K12	849	172	544	C+
College	923	139	503	C-
Immigration	693	105	767	C
Marijuana	881	118	566	C
Marriage Rights	929	105	531	B+
Reference				
Obamacare	498	-	113	B
K12	561	-	50	C
College	573	-	38	C-
Immigration	375	-	236	C+
Marijuana	498	-	113	C
Marriage Rights	554	-	57	B+

For any given issue, between 10% and 20% of those who assigned grades registered a grade change. In all, 556 out of the 1575 CRC participants changed their grades at least once (Figure 3). We also found that the aggregate results

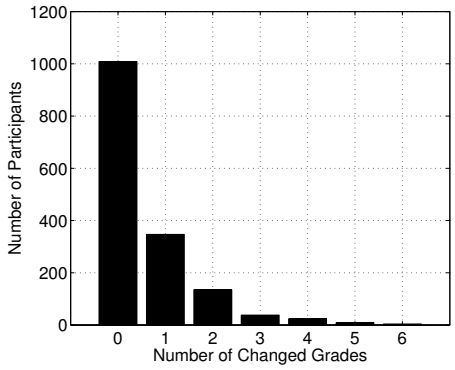

Figure 3: The majority of participants did not change grades. 35% of participants changed their grade on at least one issue, the majority of which (62%) changed only a single issue. Less than 5% of participants changed their grades on more than 4 out of the 6 issues.

of the reference survey matched the CRC quite well. On only two issues (K12 and Immigration), we found a observed differences which were both less than a letter grade (+ or -).

In our evaluation of these two surveys, we will use the unit *full letter grades*. For example, one full letter grade corresponds to the difference between an A grade and a B grade. A difference of a + or - is represented as $\frac{1}{3}$ eg. B to B+ or B+ to A-.

7.1 Social Herding Towards the Median

Using the non-parametric test proposed in Section 4, we tested the hypothesis of whether grade changes led to sig-

nificantly more concentration around the median grade. In our first experiment (Figure 4), we tested the absolute deviations of only the CRC users. We compared the group of users that did not change their grades to the group that changed their grades. We found that while there were no statistically significant differences between the initial grades of the two groups, the final grades of the group that changed were statistically significantly more concentrated than both their own initial grades and the grades of the no change group.

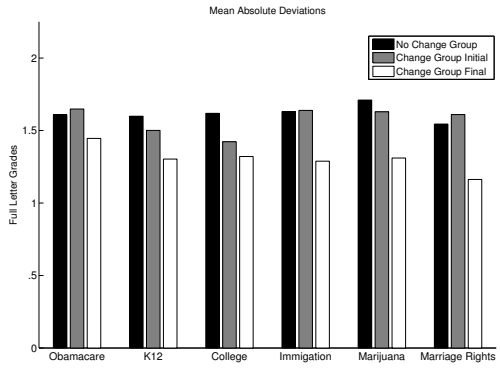

Figure 4: For those users that changed their grades, final grades were significantly more concentrated around the median grade than their initial grades. In addition, these grades are more concentrated than the grades for those who didn't change.

For the set of participants who changed their grades P_c and those who did not P_n:

Issue	p-val(P_c vs. P_n)	p-val(i vs. f)
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

These results are consistent with the social herding hypothesis. When participants change their grades, they are more likely to concentrate around the median. What is particularly surprising is that the two groups of participants P_n and P_c are very similar in terms of intial grades, and the data suggests that herding is not correlated with more or less concentrated initial grades.

In our second experiment (Figure 5), we apply the same testing procedure to compare the grades from the CRC to to those in the reference survey. We absolute deviations of the group of users who changed their grades in the CRC against users from the reference survey.

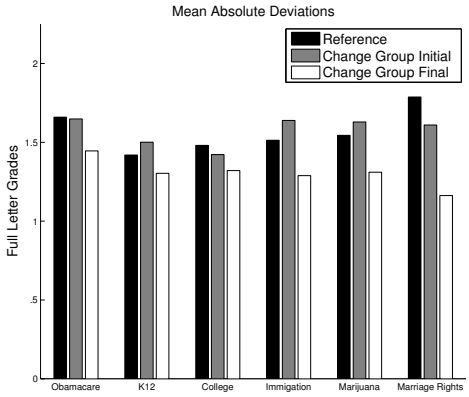

Figure 5: We found that final grades were significantly more concentrated in the CRC compared to grades in the reference survey. Similar to Figure 4, we found that there was no statistically significant difference between the reference survey and the initial grades.

Issue	p-val(R vs. i)	p-val(R vs. f)
Obamacare	0.5386	0.0015
K12	0.8283	0.0097
College	0.1452	0.0091
Immigration	0.3765	1.1787e-04
Marijuana	0.7288	9.3111e-06
Marriage Rights	0.2478	0.0161

The results of our two experiments suggest that the CRC rating data is affected by social herding. We not only found that participants' changed grades were statistically significantly more likely to concentrate around the median, they were also more likely in comparision to the reference survey. While correlation does not imply causation, we argue that this evidence is most consistent with the social herding hypothesis. As the CRC was not a randomized survey, there are possibly confounding covariates eg. participants who changed their grades were more likely to leave tightly concentrated grades in the first place. However, our comparision with the reference survey, and discovery that initial grades were largely consistent with the reference survey and with those that didn't change their grades, suggest that these confounding covariates are not very significant. These results are encouraging and we hope to run a randomized user study to confirm the causal relationship between revealing the median and concentrated grades.

7.2 Social Herding Effects

We tested the hypotheses and conclude significant additional concentration of grades around the median grade. In Section 4, we described how we could use the results of the hypothesis test to estimate the Δ parameter, which quantifies how different the hypothesis is from the null distribution. In Figure 6, we show the parameter estimates for each of the issues. As before, the units of the plot are in terms of letter grades. For the issues about Marriage Rights, we find that parameter is 2/3 of a letter grade. This means that the set of absolute deviations for the change group X_c was on aver-

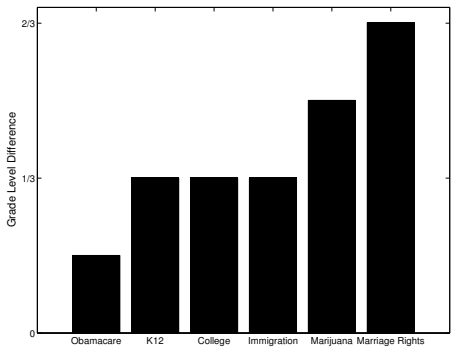

Figure 6: We calculate the parameter Δ which is the most likely amount that our observations deviate from the null hypothesis. This can be interpreted as how much more are grades in our change group concentrated around the median.

age 2/3 of a letter grade smaller. For the other issues, the parameter was smaller indicating less of an effect of social herding. The parameter was the smallest for the first issue (Obamacare), and we conjecture that for this issue many participants were still learning how to use the slider interface; leading to random grade changes. For this issue, we see that it had the most number of grade changes as well.

This parameter is very relevant to the design both recommender systems and predictive models. Consider the problem of trying to predict a particpant's next grade. The simplest model we could design is a model where we always select the median grade. Such a model is often used as a baseline for comparision in recommender systems. What we may interpret as low prediction error may in fact be the effects of social herding around the median, and if we were to apply this model asking the same questions but in a system without the herding effects; we may find that the same model performs poorly. A median prediction is a naive model, but this problem affects many recommendation algorithms since they often rely on proximity metrics such as clustering, k-nearest neighbors, and some kernel machine learning methods.

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 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 8 and Figure 9, 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,

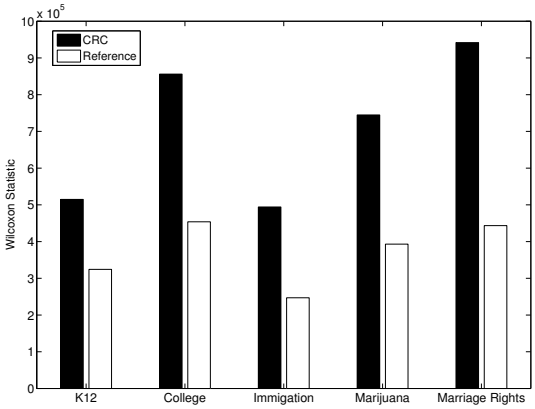

Figure 7: TODO

for Obamacare and Marriage Rights, we found that the relationship was quadratic.

Figure 9 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 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

11. REFERENCES

[1] J. Albors, J. C. Ramos, and J. L. Hervas. New learning network paradigms: Communities of objectives, crowdsourcing, wikis and open source. *International Journal of Information Management*, 28(3):194–202, 2008.

[2] S. E. Asch. Opinions and social pressure. *Readings about the social animal*, pages 17–26, 1955.

[3] S. E. Asch. *Studies of independence and conformity.* American Psychological Association, 1956.

[4] A. V. Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817, 1992.

[5] S. Bikhchandani and S. Sharma. Herd behavior in financial markets: A review. 2000.

[6] R. Bond and P. B. Smith. Culture and conformity: A meta-analysis of studies using asch's (1952b, 1956) line judgment task. *Psychological bulletin*, 119(1):111, 1996.

[7] R. E. Burnkrant and A. Cousineau. Informational and normative social influence in buyer behavior. *Journal of Consumer research*, pages 206–215, 1975.

[8] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, pages 141–150. ACM, 2009.

[9] P. M. DeMarzo, D. Vayanos, and J. Zwiebel. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3):909–968, 2003.

[10] U. M. Dholakia, S. Basuroy, and K. Soltysinski. Auction or agent (or both)? a study of moderators of the herding bias in digital auctions. *International Journal of Research in Marketing*, 19(2):115–130, 2002.

[11] R. M. Groves, F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey methodology.* John Wiley & Sons, 2013.

[12] J.-H. Huang and Y.-F. Chen. Herding in online product choice. *Psychology & Marketing*, 23(5):413–428, 2006.

[13] E. L. Lehmann and H. J. D'Abrera. *Nonparametrics: statistical methods based on ranks.* Springer New York, 2006.

[14] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025, 2011.

[15] S. Moscovici and C. Faucheux. Social influence, conformity bias, and the study of active minorities. *Advances in experimental social psychology*, 6:149–202, 1972.

[16] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[17] B. S. Noveck. Wiki-government. *Democracy: A Journal of Ideas (7)*, 2008.

[18] K. O'Hara. Transparency, open data and trust in government: Shaping the infosphere. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 223–232. ACM, 2012.

[19] B. Shneiderman. *Designing the user interface: strategies for effective human-computer interaction*, volume 2. Addison-Wesley Reading, MA, 1992.

[20] R. Sipos, A. Ghosh, and T. Joachims. Was this review helpful to you? it depends! context and voting patterns in online content.

[21] W. Wood. Attitude change: Persuasion and social influence. *Annual review of psychology*, 51(1):539–570, 2000.

[22] H. Zhu, B. Huberman, and Y. Luon. To switch or not to switch: understanding social influence in online choices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2257–2266. ACM, 2012.

APPENDIX

A. HEADINGS IN APPENDICES

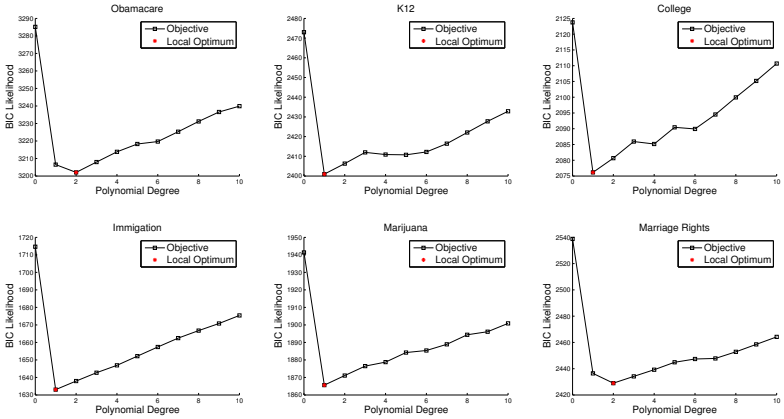

Figure 8: TODO

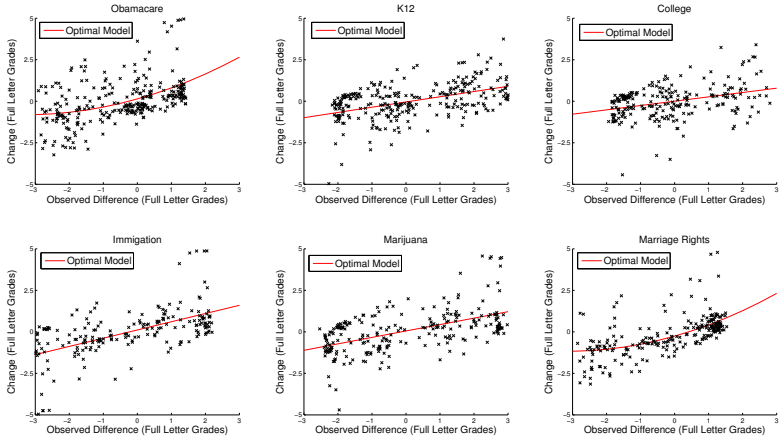

Figure 9: TODO