

**Examining the Separation-Deviation Model's Effectiveness in Group Rating Problems**

**Zoya Goel**

---

# Contents

<b>1</b>	<b>Abstract</b>	<b>2</b>
<b>2</b>	<b>Introduction</b>	<b>2</b>
<b>3</b>	<b>Literature Review</b>	<b>3</b>
<b>4</b>	<b>Simulation Setup</b>	<b>4</b>
4.1	Generating True Scores . . . . .	4
4.2	Assigning Papers to Groups . . . . .	4
4.3	Assigning Judges to groups . . . . .	5
4.4	Assigning Judges the Papers they will Judge . . . . .	5
4.5	Creating Judge Scores that include bias . . . . .	6
4.6	Reaching a Consensus Rating . . . . .	6
4.7	Can Taking the Average Score for a Paper Recover its True Score? . . . . .	7
4.8	Example . . . . .	8
<b>5</b>	<b>Experimental Results</b>	<b>8</b>
<b>6</b>	<b>Adjusting the SD Model For Scenario 2</b>	<b>11</b>
<b>7</b>	<b>When <math>g_i^{(k)}</math> is Quadratic</b>	<b>11</b>
7.1	The First Summand . . . . .	12
7.2	The Second Summand . . . . .	12
<b>8</b>	<b>Discussion and Future Work</b>	<b>13</b>
<b>9</b>	<b>Acknowledgements</b>	<b>14</b>
	<b>References</b>	<b>15</b>

## 1 Abstract

Group rating problems are relevant when considering decision-making processes where individuals have to be evaluated by judges. We focus on a potential method for solving consensus rating problems - the Separation-Deviation (SD) model. This model considers a set of ratings given to individuals (which may be complete or incomplete) and returns a final score for each individual. We carry out experiments to determine how effective the SD model is at retrieving the “true” scores of each individual, and we compare its performance to a simple heuristic that finds the mean scores of each individual. We find that the SD model generally does not outperform this heuristic, regardless of whether judges give individuals biased or unbiased ratings. We also find that when we choose different convex functions for different individuals used in the SD model to correct for bias, we can improve the SD model’s performance for minority groups that face bias. More work needs to be done applying the SD model to “true” score distributions that aren’t just the normal distribution, and in situations where bias is applied more realistically.

## 2 Introduction

Finding a consensus rating or ranking is a common problem in many contexts, from judging papers in a paper competition [2], to hiring individuals for work [3] [6]. Additionally, bias tends to infiltrate these processes [3] [6]. Discrimination and bias in the workplace have been extensively documented, and this bias tends to change in nature and intensity depending on gender and race [7]. There have been many examples of gender bias observed by researchers, wherein employers tend to favor male job applicants over equally qualified female ones[7]. Researchers have found that humans tend to rely on their biases more as they have to make decisions with less information [4]. This study helps explain why bias runs rampant in the hiring process, as hiring committees have to determine how effective a prospective employee might be based on an incredibly limited amount of information. Furthermore, this information is often far from objective - it has been shown time and time again that job interviews tend to hurt minority candidates due to the personal biases of the interviewers [5].

All of these issues that come with rating or ranking entities necessitate interventions in order to counteract them. Suhr et al. and Peng et al. do this at the beginning of the ranking process. In both of these papers, participants rank sets of male and female candidates in order of preference for specific tasks. Both studies attempt to rectify gender bias by presenting the candidates to the participants in different ways. In Suhr et al, the candidates are shown in a ranking generated from one of three possible methods, one of which was a fair ranking algorithm. In Peng et al, they changed the proportion of males and females in the slates of candidates presented to participants. Both of these interventions were sometimes successful, but there were some contexts where gender biases were more persistent, which is where these interventions faltered. Both of these studies, while demonstrating the effectiveness of certain interventions, also demonstrate the persistent nature of bias in some situations.

The intervention that this report focuses on, the Separation-Deviation (SD) model [1] [2], is one that occurs later in the process - it intends to find a consensus rating amongst many ratings provided by reviewers. This intervention also functions in a situation where the set of ratings is incomplete - in other words, when each reviewer reviews a subset of entities, rather than all of them at once. This model generates a consensus rating by turning it into a minimization problem.

They assign penalties for differences in scores between the consensus rating and each judge rating, as well as separation gaps - the difference in score between two papers in the same rating. Once these penalties are minimized, the model yields a consensus rating.

In this report, we demonstrate the effectiveness of this solution by creating a simulation that mimics a possible situation where a consensus rating must be reached. Additionally, we compare it to perhaps the most straightforward way of creating a consensus rating - finding the mean score for each object. We also find a way to pick an optimal amount of reviewers for each object if we want to find a consensus rating using this method.

### 3 Literature Review

In our day and age, bias in the workplace is a problem that affects many women and minorities, especially in fields related to Science, Technology, Engineering, and Mathematics. This gender bias has an enormous impact on how women function and carry themselves in a work environment. Researchers have also documented how this bias changes in intensity and nature depending on race as well. Naturally, when there is bias in the workplace, it will extend to hiring processes as well. Studies have shown that people are twice as likely to hire a man instead of a woman for a job that requires mathematics knowledge, even when their math skills are identical. Additionally, faculty at research universities, when reviewing applications of hypothetical applicants, were more likely to view a man as “more competent and hireable” when both applicants have the same applications [7]. Lastly, researchers have found that humans tend to rely on their biases more as they have to make decisions with less information [4]. This study helps explain why bias runs rampant in the hiring process, as hiring committees have to determine how effective a prospective employee might be based on an incredibly limited amount of information. Furthermore, this information is often far from objective - it has been shown time and time again that job interviews tend to hurt minority candidates due to the personal biases of the interviewers [5].

In Suhr et. al. [6], researchers seek to answer three questions: (1) Do employers exhibit gender bias uniformly across different contexts? (2) How effective are fair ranking algorithms at remedying gender bias? (3) Can fair ranking algorithms lead to disparate outcomes for different underrepresented groups? They tested this by using Amazon MTurk workers as proxy employers and showing them one of three possible rankings they could make decisions based off of. One method was generating the ranking by using the default recommendation algorithm from the job search website they used to find candidates, TaskRabbit.com. Another method was using FairDetGreedy, a fair ranking algorithm applied as a post-processing step to TaskRabbit’s recommendation algorithm. The last possible method was that the candidates were presented in a random ranking. They found that fair ranking algorithms can help remedy some gender bias, but in certain jobs where employers tend to have a male preference, they are less effective. They also found that participants in their experiment tended to try and correct their self-perceived biases [6]

Another paper that does this is a paper by Peng et. al. [3]. This paper aims to see if, when displaying candidates to a hypothetical employer, changing the proportions of genders on each slate of candidates helps mitigate bias. They found that this works for most professions, but there are some professions where there is so much bias that this mitigation technique is not enough. Additionally, researchers also found that the personal features of the decision-makers can impact outcomes. This paper is similar to the previous paper [6], because it arrives at similar results -

despite using a bias mitigation technique, there are some professions that people tend to have a strong gender bias towards. This paper also uses a similar methodology as Sühr et al - they use Amazon MTurk in order to gain participants for their experiment and structure their trials in a somewhat similar way. (Peng et al.) [3].

In Hochbaum et al. [1], the authors introduce the Separation-Deviation (SD) model, and in Hochbaum et al. [2], the authors apply the SD model to a student paper competition. The SD model is used for aggregate rating scenarios. For each object being reviewed (in this case, papers), point-wise scores and separation gaps between papers are taken as inputs to penalty functions. Penalties are incurred when the scores and separation gaps in the aggregate rating differ from those of the judges’ ratings. The goal is to minimize these penalties, which can be done in polynomial time if the penalty functions are convex. Penalizing differences in both separation gaps and scores might seem redundant, but the authors’ motivation for doing this is to indirectly penalize a change in ranking, even if only scores are provided. In the work done by Hochbaum et al., they apply the SD model to a paper competition by modifying the penalty functions to be more appropriate for what data they have. They then use the competition data and different versions of the model they created to create aggregate ratings and rankings. They justify decisions for certain papers that the SD model arrives at, but they don’t necessarily compare the model’s effectiveness to other possible mechanisms for determining aggregate ratings.

## 4 Simulation Setup

### 4.1 Generating True Scores

We made this simulation in the context of Hochbaum et al.[2], which used data from a paper competition. So, we refer to our reviewers as judges, and the objects being reviewed as papers. Utilizing the notation from (Hochbaum 2021), we are given a set of  $V$  papers of size  $n$ , and we assign a unique identifier to each element such that  $V = \{1, 2, \dots, n\}$ . First, we define a vector which contains true scores for each of the papers.

**Definition 4.1** (True Score Vector). Let  $b$  be a vector of size  $n$ . Entry  $b_i$  corresponds to the true score of paper  $i$ , where  $i \in [1, n]$ . We pull the value of  $b_i$  from some normal distribution,  $b_i \sim N(\mu, \sigma_1)$  - this represents the distribution of true scores, with some mean  $\mu \in [u, \ell]$ , where  $u$  and  $\ell$  are the upper and lower score limits respectively, and some standard deviation  $\sigma_1$ . If we draw a  $b_i > u$  or a  $b_i < \ell$ , we make  $b_i = u$  or  $b_i = \ell$ , respectively.

### 4.2 Assigning Papers to Groups

When we assign papers to groups, this could translate to protected classes in the real world (ex: papers in Group 1 were written by men, and papers in Group 2 were written by women). We define our possible paper groups as follows.

**Definition 4.2** (Paper Groups). Let our set of groups  $G$ , be defined as  $G = \{G_1, G_2, \dots, G_{|G|}\}$ , where  $|G| \in [1, n]$ .

We then define each paper’s group number as follows.

**Definition 4.3** (Paper Group Numbers). Let vector  $g$  be a vector of size  $n$  containing the group numbers of each paper. Let  $g_i$ , where  $i \in [1, n]$  be paper  $i$ ’s group number - in other words, paper  $i$  is in  $G_{g_i}$ . For each  $g_i$ , we pull the value from a categorical distribution with  $|G|$  events - if it equals 1 for some event  $\alpha \in [1, |G|]$ , this means that  $g_i = \alpha$ , so paper  $i$  is in  $G_\alpha$ .

We then define the conditions of the categorical distribution used to determine each  $g_i$ .

**Definition 4.4** (Categorical Distribution Conditions for Paper Groups). Let  $p_{G_\alpha}$  be the proportion of papers we aim to assign to  $G_\alpha$ . For the categorical distribution used to determine  $g_i$ ,  $\sum_{\alpha=1}^{|G|} p_{G_\alpha} = 1$ .

### 4.3 Assigning Judges to groups

When we assign judges to groups, this could translate to judges with different sets of biases (ex: judges from Group 1 tend to rate women more poorly than men, judges from Group 2 tend to be closer to true scores). The process of assigning judges to groups is essentially the same as the process for assigning papers to groups. Utilizing the notation from (Hochbaum, 2021), assume we have  $K$  judges, with each judge  $k \in 1, 2, \dots, K$ . We define our set of groups as follows.

**Definition 4.5** (Judge Groups). Let our set of judge groups,  $H$ , be defined as  $H = \{H_1, H_2, \dots, H_{|H|}\}$ , where  $|H| \in [1, K]$ . We then define the judge groups as follows.

We then define each judge’s group number as follows

**Definition 4.6** (Judge Group Numbers). Let vector  $h$  with size  $K$  contain the group numbers for each of the judges. Let  $h_k$ , where  $k \in [1, K]$  be judge  $k$ ’s group number - in other words, judge  $k$  is in  $H_{h_k}$ . For each  $h_k$ , we pull the value from a categorical distribution with  $|H|$  events - if it equals 1 for some event  $\beta \in [1, |H|]$ , this means that  $h_k = \beta$ , so judge  $k$  is in  $H_\beta$ .

We then define the conditions of the categorical distribution used to determine each  $h_k$ .

**Definition 4.7** (Categorical Distribution Conditions for Judge Groups). Let  $p_{H_\beta}$  be the proportion of judges we aim to assign to  $H_\beta$ . For the categorical distribution used to determine  $h_k$ ,  $\sum_{\beta=1}^{|H|} p_{H_\beta} = 1$ .

### 4.4 Assigning Judges the Papers they will Judge

Utilizing the notation from (Hochbaum 2021), we define a set of scores, or ratings, that a judge  $k$  provides for each paper as follows.

**Definition 4.8** (Judge Rating Vector). We define the papers that judge  $k$  judges as a vector  $a^{(k)}$  with size  $n$ . Thus  $a_i^{(k)}$  corresponds to judge  $k$ ’s rating of paper  $i$ . Any papers that a judge does not review have their corresponding entries assigned as undefined.

Assume we would like to assign each judge  $f$  papers to judge, where  $f \in [1, n]$ . The method used for doing this is that we essentially assign judge 1 the first  $f$  papers, judge 2 the next  $f$  papers, and so on. If we reach judge  $i$  and there are less than  $f$  papers left to judge, we give the judge those last papers and begin assigning at the first paper again. We formalize this as follows.

**Theorem 4.1** (Paper Assignment Method). *Vector  $a^{(k)}, a_i^{(k)}$  is defined for all  $i \in [(k * f - (f - 1)) \bmod n, k * f \bmod n]$ , and all other entries are undefined.*

Furthermore, we can use  $f$  to find the number of reviews per paper.

**Theorem 4.2** (Reviewers Per Paper). *Given that  $f$  is the number of papers assigned to each judge,  $K$  is the total number of judges, and  $n$  is the total number of papers, the number of reviews per paper,  $r = Kf/n$*

#### 4.5 Creating Judge Scores that include bias

As mentioned previously, we could have groups of judges with different sets of biases. This might cause them to rate papers from different groups differently. For example, a judge from Group 1 might rate papers from Group 1 with some bias from the true score, but papers from Group 2 are rated similarly to their true scores. A judge from Group 2 might have different biases for both groups. To represent this, we will add a bias quantity based on the paper group and the judge group to the true paper score, which will then become the judge's score for that particular paper. We define the bias quantity as follows.

**Definition 4.9** (Bias Quantity). For any possible pairwise combination of paper and judge groups  $(\alpha, \beta)$ , let  $D_{\alpha, \beta}$  represent the bias a judge from group  $\beta$  adds onto a paper from group  $\alpha$ , where  $D_{\alpha, \beta} \sim N(0, \sigma_{\alpha, \beta})$ .

Then, we define the value of an entry in  $a^{(k)}$  with respect to the true score and the bias quantity.

**Definition 4.10** (Judge Score of Paper). Recall that  $a_i^{(k)}$  is the score of paper  $i$  judged by judge  $k$ . We define  $a_i^{(k)} = b_i + D_{g_i, h_k}$ . However, if  $b_i + D_{g_i, h_k} < \ell$ , then  $a_i^{(k)} = \ell$ . If  $b_i + D_{g_i, h_k} > u$ , then  $a_i^{(k)} = u$

Note that every time we have the same judge group and paper group combination, we draw a new bias quantity from its distribution every time.

#### 4.6 Reaching a Consensus Rating

Using (Hochbaum 2006) and (Hochbaum 2021), we define the SD model. First, we must define the separation gap:

**Definition 4.11** (Separation Gap).

$$p_{ij} = \begin{cases} a_i^{(k)} - a_j^{(k)} & \text{if } i, j \in \mathcal{A}^{(k)} \\ \text{undefined} & \text{otherwise} \end{cases}$$

We define  $\mathcal{A}^{(k)}$  as the set of papers that judge  $k$  judges.

**Definition 4.12** (SD Model). If we have a consensus rating, represented as a vector  $x$ , where entry  $x_i$  represents paper  $i$ 's score in the consensus rating, we aim to minimize the following with respect to  $x$  and  $z$ :

$$\sum_{k=1}^K \sum_{i=1}^n \sum_{j=1}^n f_{ij}^{(k)} (z_{ij} - p_{ij}) + \sum_{k=1}^K \sum_{i=1}^n g_i^{(k)} (x_i - a_i^{(k)})$$

subject to:

- a.  $z_{ij} = x_i - x_j, i, j \in [1, n]$
- b.  $x_i \in [\ell, u]$
- c.  $x_i \in \mathbb{Z}$

Functions  $f$  and  $g$  are penalty functions and must be convex. Additionally, if any  $a_i^{(k)}$  or  $a_j^{(k)}$  is undefined (in other words, a judge hasn't judged that paper), the functions evaluate to 0, so a penalty isn't added.

#### 4.7 Can Taking the Average Score for a Paper Recover its True Score?

Perhaps the most straightforward way to attempt to recover the true score is to take the mean of judge scores and assign that as the true score. Because of this, we may want to figure out an upper bound for how many reviews a paper should have to recover the true score within a certain accuracy. We can use the Hoeffding Inequality for independent sub-Gaussian random variables to solve this:

**Definition 4.13** (Hoeffding Inequality for independent sub-Gaussian random variables). Let  $X_1 \dots X_m$  be independent centered random variables where for each  $j \in [1, m]$ ,  $X_j$  is sub-Gaussian with upper bound on its variance  $v_j \in (0, \infty)$  and let  $(c_1, \dots, c_m) \in \mathcal{R}^n$ . Let  $S_n = \sum_{j \leq m} c_j X_j$ . For all  $\beta > 0$ ,

$$P(|S_n| \geq \gamma) \leq \exp\left(-\frac{\gamma^2}{2 \sum_{j=1}^m c_j v_j}\right)$$

If we strategically substitute variables used in the simulation into the Hoeffding inequality, we can acquire an inequality that we can use to figure out how many reviews a paper needs.

**Proposition 4.3.** *Let  $\mathcal{I}$  be the set of judges judging paper  $i$ , and let  $|\mathcal{I}| = K_i$ . Using the above theorem, let  $S_n = \sum_{k \in \mathcal{I}} a_i^{(k)} - K_i b_i$ . Then, let  $\gamma = K_i t$ , and let  $c_i = 1$ . Also, assume that all variables have the same upper bound on variance,  $v$ . This means that:*

$$P(|S_n| \geq \gamma) = P(\bar{a}_i - b_i \geq t) \leq \exp\left(-\frac{K_i t^2}{2v}\right)$$

, where value  $\bar{a}_i$  is the average judge score for paper  $i$ , and  $t$  is our chosen threshold for  $\bar{a}_i - b_i$ . However, we will consider the absolute value of  $\bar{a}_i - b_i$ , so

$$P(|\bar{a}_i - b_i| \geq t) \leq 2 \exp\left(-\frac{K_i t^2}{2v}\right)$$

Since the difference between any judge score and the true score will be the bias quantity,  $v$  will be the upper bound of the bias quantity. Recall that  $D_{\alpha, \beta} \sim N(0, \sigma_{\alpha, \beta})$ . So, we can approximate  $v = z \sigma_{\alpha, \beta}$ , where  $z$  is the  $z$  value for our chosen (ideally large) percentile (ex: 99th percentile) from the standard normal distribution. We also may choose to round up or down depending on our situation. From this, we can derive our inequality for the minimum required judges for judging paper  $i$ .

**Definition 4.14** (Inequality for Minimum Number of Reviews for Paper  $i$ ).

$$K_i \geq \ln\left(\frac{2}{P(|\bar{a}_i - b_i| \geq t)}\right) \frac{2v}{t^2}$$

### 4.8 Example

Assume we want the difference between average judge score and true score, to be less than 1, with a probability of 95%. This means  $P(|\bar{a}_i - b_i| < t = 1) = 0.95$ , so  $1 - P(|\bar{a}_i - b_i| < t = 1) = P(|\bar{a}_i - b_i| \geq t) = 0.05$ . Additionally, assume  $\sigma_{\alpha, \beta} = 2$ , and assume the largest possible bias is less than or equal to the value in the 99th percentile for  $N(0, 2)$ . This means that  $v = 4.66$ , but we will round up (we want integer scores) so  $v = 5$ . This means that:

$$K_i \geq \ln\left(\frac{2}{0.05}\right) \frac{2(5)}{1^2} \approx 36.8879454$$

This means that we need at least 37 reviews for paper  $i$  for the difference between the mean score and judge score to be less than 1, with a probability of 95%. Additionally, if we define the number of reviewers a paper has as  $r$ , if we use Theorem 3.2,  $f = rn/K$ . Since we now have a number for  $r$ , we can figure out what value we need for  $f$  to make sure each paper gets at least  $r$  amount of reviews, which, if we have 58 papers and 63 judges like in (Hochbaum 2021),  $f \approx 33.96111243$ , which rounds up to assigning 34 papers per judge.

## 5 Experimental Results

To compare the accuracy of the SD Model and a simple heuristic that just computes the mean score to generate a consensus rating, we calculate the average percentage error in scores from both these methods. Each scenario is run 40 times, for 60 papers and 60 judges (similar to the 58 papers and 63 judges in Hochbaum et. al [2]). We consider two groups of papers, and two groups of judges ( $|G| = 2, |H| = 2$ ). We assign papers to judges from each of the groups uniformly at random. We include a plot (Figure 1) that shows the number of reviews that occur per every possible author-judge group pairing. We also slightly relax the integer constraint when using the SD model to two decimal points instead. This also means that any true scores and judge scores are truncated to two decimal points.

We carry out this entire process four times for each scenario and  $f$  value, but each time we run this process we use a different objective function for the SD model. More specifically, we change the  $g_i^{(k)}$  function. There are four possible  $g_i^{(k)}$  functions that we use:

1.  $g_i^{(k)}(x_i - a_i^{(k)}) = (x_i - a_i^{(k)})^2$
2.  $g_i^{(k)}(x_i - a_i^{(k)}) = \exp(-1(x_i - a_i^{(k)})) * \lambda$ , where  $\lambda = 0.1$
3.  $g_i^{(k)}(x_i - a_i^{(k)}) = \exp(-1(x_i - a_i^{(k)})) * \lambda$ , where  $\lambda = 0.05$
4.  $g_i^{(k)}(x_i - a_i^{(k)}) = \exp(-1(x_i - a_i^{(k)})) * \lambda$ , where  $\lambda = 0.01$

- **Scenario 1.** We first consider the case where each judge is assigned 5 papers ( $f = 5$ ) and compare the errors to the case when each judge is assigned 20 papers each ( $f = 20$ ). Here’s the data as we generate it:

1. **True scores:**  $b_i \sim \max(1, \min(N(\mu = 5, \sigma_1 = 2), 10))$  for each paper  $i$ .
2. Each paper is assigned to Group 1 or Group 2 uniformly at random. Each judge is assigned Group 1 or Group 2 also uniformly at random.
3. **Judges scores for papers:**
  - Each judge scores each paper as  $a_i^{(k)} = \max(1, \min(b_i + D_{\alpha,\beta}, 10))$ , where  $D_{\alpha,\beta} \sim N(0, 1.5)$ .
4. Recall that the separation gap for papers  $i$  and  $j$  as judged by judge  $k$  is  $p_{ij} = a_i^{(k)} - a_j^{(k)}$ . Also recall that the separation gap for papers  $i$  and  $j$  in the consensus rating is  $z_{ij} = x_i - x_j$ . The penalty function,  $f_{ij}^{(k)}(z_{ij} - p_{ij}) = (z_{ij} - p_{ij})^2$

We collect the percent error for all reviews across all trials for this scenario and compile them into box plots (Figure 2, Figure 3). We calculate percentage error in an experiment as follows:  $(x_i - b_i)/b_i$ . This should be small as the number of judges per paper increases. We also make histograms comparing the output of each method and the true scores (Figure 4, Figure 5). Furthermore, we ran a one-way ANOVA to compare all of these outputs and found that  $p = 0.00 < 0.05$  when  $f = 5$ . As a result, we ran Tukey’s Honest Significant Difference (HSD) test to compare every method to each other. From running this test, we found that the results from when the SD model’s  $g_i^{(k)}$  function was quadratic and the results from taking the mean were not significantly different ( $p = 1.00 > 0.05$ ). Other than this pair, all of the other methods performed significantly differently from each other ( $p = 0.00 < 0.05$ ). When comparing each of these methods to the true scores, we found that the results from when the SD model’s  $g_i^{(k)}$  function was quadratic and the results from taking the mean were not significantly different from the true scores ( $p = 0.989 > 0.05$ ). The rest of the methods performed significantly differently from the true scores ( $p = 0.00 < 0.05$ ). We get somewhat similar results when  $f = 20$  as well. We ran a one-way ANOVA to compare all of these outputs and found that  $p = 0.00 < 0.05$ . As a result, we ran Tukey’s HSD. From running this test, we found that the results from when the SD model’s  $g_i^{(k)}$  function was quadratic and the results from taking the mean were not significantly different ( $p = 1.00 > 0.05$ ). We also found that the results from the SD model when  $\lambda = 0.1$  and when  $\lambda = 0.05$  were not significantly different ( $p = 0.865 > 0.05$ ), and the results from the SD model when  $\lambda = 0.01$  and when  $\lambda = 0.05$  were also not significantly different ( $p = 0.320 > 0.05$ ). However, the results from the SD model when  $\lambda = 0.1$  and when  $\lambda = 0.01$  were significantly different ( $p = 0.018 < 0.05$ ). All other possible pairs of methods also performed significantly differently ( $p = 0.00 < 0.05$ ). When comparing each of these methods to the true scores, we found that the results from when the SD model’s  $g_i^{(k)}$  function was quadratic and the results from taking the mean were not significantly different from the true scores ( $p = 1.000 > 0.05$ ). The rest of the methods performed significantly differently from the true scores ( $p = 0.00 < 0.05$ ).

- **Scenario 2.** This scenario is almost identical to the first one, but we simulate “favoritism,” where if the judges judge a paper from the same group that they are from, the bias is more likely to be positive. We also simulate having a minority group, Group 2. Like the first scenario, we consider the case where each judge is assigned 5 papers ( $f = 5$ ) and compare the errors to the case when each judge is assigned 20 papers each ( $f = 20$ ).

1. **True scores:**  $b_i \sim \max(1, \min(N(\mu = 5, \sigma_1 = 2), 10))$  for each paper  $i$ .
2. Approximately 4/5 of the judges are in Group 1, and 1/5 are in Group 2. Recall our method for assigning judge groups in Definitions 3.6 and 3.7. We have a categorical distribution with 2 events since  $|H| = 2$ , where the probability of a judge being in

Group 1 is 4/5, and the probability of a judge being in Group 2 is 1/5. We use the exact same categorical distribution to find the paper groups since  $|G| = 2$ . We include a plot (Figure 6) that shows the number of reviews that occur per every possible author-judge group pairing for this scenario.

3. **Judges' in group  $\beta$  scores for papers in group  $\alpha$  when  $\alpha = \beta$ :** assuming judges rate papers from their groups better, compared to papers from other groups.
  - If judge  $k$  is in group  $s$  ( $\beta = 1$ ) and evaluating paper  $i$  in group  $s$  ( $\alpha = 1$ ), then score  $a_i^{(k)} = \max(1, \min(b_i + D_{\alpha,\beta}, 10))$ , where  $D_{\alpha,\beta} \sim N(0, 2)$ .
4. **Judges' in group  $\beta$  scores for papers in group  $\alpha$  when  $\alpha \neq \beta$ :**
  - If judge  $k$  is in group  $\beta$  and evaluating paper  $i$  in group  $\alpha$  ( $\alpha \neq \beta$ ), then score  $a_i^{(k)} = \max(1, \min(b_i - 2 + D_{\alpha,\beta}, 10))$ , where  $D_{\alpha,\beta} \sim N(0, 2)$
5. Recall that the separation gap for papers  $i$  and  $j$  as judged by judge  $k$  is  $p_{ij} = a_i^{(k)} - a_j^{(k)}$ . Also, recall that the separation gap for papers  $i$  and  $j$  in the consensus rating is  $z_{ij} = x_i - x_j$ . The penalty function,  $f_{ij}^{(k)}(z_{ij} - p_{ij}) = (z_{ij} - p_{ij})^2$

We collect the percent error for all reviews across all trials for this scenario and compile them into box plots (Figure 7, Figure 8). We calculate the error in the same way we did in Scenario 1. We also make histograms comparing the output of each method and the true scores (Figure 9, Figure 10). Furthermore, we ran a one-way ANOVA to compare all of these outputs. When running ANOVA for the Group 1 output when  $f = 5$ ,  $p = 0.00 < 0.05$ . As a result, we run Tukey's HSD. When running this test, we find that we found that the results from when the SD model's  $g_i^{(k)}$  function was quadratic and the results from taking the mean were not significantly different from each other ( $p = 1.000 > 0.05$ ). However, every other possible pair of methods was significantly different from each other ( $p = 0.00 < 0.05$ ). Also, every method was significantly different from the true scores ( $p = 0.00 < 0.05$ ). The results for the one-way ANOVA and Tukey's HSD tests performed on Group 2 when  $f = 5$  were exactly the same as Group 1's, except that when running Tukey's HSD, the results from the SD model when  $\lambda = 0.1$  and when  $\lambda = 0.05$  were not significantly different ( $p = 0.141 > 0.05$ ). We get somewhat similar results for Group 1 when  $f = 20$  as well. We ran a one-way ANOVA to compare all of these outputs and found that  $p = 0.00 < 0.05$ . As a result, we ran Tukey's HSD. From running this test, we found that the results from when the SD model's  $g_i^{(k)}$  function was quadratic and the results from taking the mean were not significantly different ( $p = 1.00 > 0.05$ ). We also found that the results from the SD model when  $\lambda = 0.1$  and when  $\lambda = 0.05$  were not significantly different ( $p = 0.515 > 0.05$ ) However, the results from the SD model when  $\lambda = 0.1$  and when  $\lambda = 0.01$  were significantly different ( $p = 0.035 < 0.05$ ). All other possible pairs of methods also performed significantly differently ( $p = 0.00 < 0.05$ ). When comparing each of these methods to the true scores, we found that all of the methods performed significantly differently from the true scores ( $p = 0.00 < 0.05$ ). For Group 2, we ran a one-way ANOVA to compare all of these outputs and found that  $p = 0.00 < 0.05$ . As a result, we ran Tukey's HSD. From running this test, we found that the results from when the SD model's  $g_i^{(k)}$  function was quadratic and the results from taking the mean were not significantly different ( $p = 1.00 > 0.05$ ). Also, the results from the SD model when  $\lambda = 0.1$  and when  $\lambda = 0.05$  were not significantly different ( $p = 0.619 > 0.05$ ), as well as when  $\lambda = 0.01$  and  $\lambda = 0.1$  ( $p = 0.197 > 0.05$ ), and  $\lambda = 0.01$  and  $\lambda = 0.005$  ( $p = 0.979 > 0.05$ ). When comparing these methods to the true scores, every method was significantly different ( $p = 0.00 < 0.05$ ), except for the SD model when  $\lambda = 0.05$  ( $p = 0.082 > 0.05$ ) and when  $\lambda = 0.01$  ( $p = 0.374 > 0.05$ ).

## 6 Adjusting the SD Model For Scenario 2

The SD model when  $g_i^{(k)}$  is quadratic and the mean tends to have the lowest error rates across the board for all scenarios (Figure 2, Figure 3, Figure 7, Figure 8). However, in Scenario 2, the SD model when  $\lambda = 0.01$  appears to perform better in terms of percent error for Group 2 both when  $f = 5$  and  $f = 20$  7 8. So, we attempt to outperform the mean and the SD model when  $g_i^{(k)}$  is quadratic by modifying the SD model as follows for Scenario 2:

**Proposition 6.1** (Modified SD Model for Scenario 2). *Consider the function  $g_i^{(k)}(x_i - a_i)$  used in the SD model. Recall that  $g_i$  is the paper group for paper  $i$ , and recall that in our simulation,  $g_i = 1$  or  $g_i = 2$ . If  $g_i = 1$ ,  $g_i^{(k)}(x_i - a_i) = (x_i - a_i)^2$ . If  $g_i = 2$ ,  $g_i^{(k)}(x_i - a_i) = \exp(x_i - a_i) * \lambda$ , where  $\lambda = 0.01$ .*

We compared this model to the SD model when  $g_i^{(k)}$  is quadratic and taking the mean by only running Scenario 2. Like the previous experiments, this scenario is run 40 times, for 60 papers and 60 judges. We consider two groups of papers, and two groups of judges ( $|G| = 2, |H| = 2$ ). We collect the percent error for all reviews across all trials for this scenario and compile them into box plots (Figure 11, Figure 12). We also create histograms that compare the output of each method and the true scores (Figure 13, Figure 14). We also run a one-way ANOVA test for outputs when  $g_i = 1, f = 5, g_i = 2, f = 5, g_i = 1, f = 20$ , and when  $g_i = 2, f = 20$ . In all situations, the one-way ANOVA yielded  $p = 0.00 < 0.05$ . As a result, we ran Tukey's HSD. In all situations, taking the mean was not significantly different from the SD model when  $g_i^{(k)}$  is quadratic ( $p = 1.000 > 0.05$ ), but the adjusted SD model was significantly different from both of these methods ( $p = 0.000 < 0.05$ ). All methods were significantly different from the true scores ( $p = 0.000 < 0.05$ ) in all cases, except for the modified SD model, which was not significantly different from the Group 2 true scores when  $f = 20$  ( $p = 0.964 > 0.05$ ). From this, we can see that when we adjust the  $g_i^{(k)}$  function used in the SD model depending on paper group membership, we can outperform the results from taking the mean and SD model when  $g_i^{(k)}$  is always quadratic, which were the models that had the lowest percent errors in most cases (Figure 7, Figure 8, Figure 11, Figure 12).

## 7 When $g_i^{(k)}$ is Quadratic

From the experimental results section, we can see that when  $g_i^{(k)}(x_i - a_i) = (x_i - a_i)^2$ , the SD model's output does not differ significantly from the output from taking the mean judge score for each paper. In this section, we present some reasoning as to why that is the case.

Our goal is to minimize the objective function with respect to  $x_i$  - some paper score in the final consensus rating yielded by the SD model. So, we will take the derivative of the objective function, set it equal to 0, and solve for  $x_i$ . We will consider each summand of the objective function one at a time.

## 7.1 The First Summand

Recall that the first summand of the SD model's objective function is:

$$\sum_{k=1}^K \sum_{i=1}^n \sum_{j=1}^n f_{ij}^{(k)}(z_{ij} - p_{ij})$$

Also, recall that:

$$p_{ij} = \begin{cases} a_i^{(k)} - a_j^{(k)} & \text{if } i, j \in \mathcal{A}^{(k)} \\ \text{undefined} & \text{otherwise} \end{cases}$$

Where we define  $\mathcal{A}^{(k)}$  as the set of papers that judge  $k$  judges. Also,  $z_{ij} = x_i - x_j$ , and  $f_{ij}^{(k)}(z_{ij} - p_{ij}) = (z_{ij} - p_{ij})^2$ . We will define the set of judges that judge paper  $i$  as  $k_1, k_2, \dots, k_q$ . So, when we fully write out this part of the objective function, we will have:

$$(x_i - x_j - a_i^{(k_1)} + a_j^{(k_1)})^2 + (x_i - x_j - a_i^{(k_2)} + a_j^{(k_2)})^2 + \dots + (x_i - x_j - a_i^{(k_q)} + a_j^{(k_q)})^2$$

However, while we have considered  $f_{ij}^{(k)}(z_{ij} - p_{ij})$ , we must also consider the opposite case,  $f_{ji}^{(k)}(z_{ji} - p_{ji})$ . So, we will also add the following terms to the terms above:

$$(x_j - x_i - a_j^{(k_1)} + a_i^{(k_1)})^2 + (x_j - x_i - a_j^{(k_2)} + a_i^{(k_2)})^2 + \dots + (x_j - x_i - a_j^{(k_q)} + a_i^{(k_q)})^2$$

We repeat this process for all possible  $j$  papers with which paper  $i$  can be compared. We will also repeat this process for other pairs of papers to compare that do not include paper  $i$ . However, those terms become irrelevant, as when we take the derivative of the entire objective function, they become 0 because we are taking the derivative with respect to  $x_i$ . So, we end up with this:

$$\begin{aligned} & \sum_{j=1}^n 2x_i - 2x_j - 2a_i^{(k_1)} + 2a_j^{(k_1)} + 2x_i - 2x_j - 2a_i^{(k_2)} + 2a_j^{(k_2)} + \dots + 2x_i - 2x_j - 2a_i^{(k_q)} + 2a_j^{(k_q)} \\ & + 2x_j - 2x_i - 2a_j^{(k_1)} + 2a_i^{(k_1)} + 2x_j - 2x_i - 2a_j^{(k_2)} + 2a_i^{(k_2)} + \dots + 2x_j - 2x_i - 2a_j^{(k_q)} + 2a_i^{(k_q)} \end{aligned}$$

We can see that everything inside this summand will end up being 0. Because of this, when we minimize the objective function with respect to  $x_i$ , we only have to consider the effect of the second summand of the objective function.

## 7.2 The Second Summand

Recall that the second summand of the SD model's objective function is:

$$\sum_{k=1}^K \sum_{i=1}^n g_i^{(k)}(x_i - a_i^{(k)})$$

We've previously stated that  $g_i^{(k)}(x_i - a_i^{(k)}) = (x_i - a_i^{(k)})^2$ . Taking the derivative of this with respect

to  $x_i$  yields:

$$2(x_i - a_i^{k_1}) + 2(x_i - a_i^{k_2}) + \dots + 2(x_i - a_i^{k_q})$$

Finally, we can define the equation that we need to solve to find the value of  $x_i$  that minimizes the objective function:

$$0 = 2qx_i - 2a_i^{k_1} - 2a_i^{k_2} + \dots + 2a_i^{k_q}$$

Then, we solve the equation:

$$x_i = \frac{a_i^{k_1} + a_i^{k_2} + \dots + a_i^{k_q}}{q}$$

Notice that this essentially means that to minimize the objective function with respect to  $x_i$ , we must take the average of all the judge scores. Therefore this finding might perhaps explain why the SD model outputs are incredibly similar to the outputs resulting from taking the average judge score for each paper. However, while the results from the SD model and the mean are very similar, they can vary slightly from each other, but not incredibly significantly (Figure 15). This difference can be explained by the fact that they are random samples.

## 8 Discussion and Future Work

From our results, we can see that neither the Separation-Deviation model nor taking the mean outperforms each other in the consensus rating problem, even in cases of favoritism/bias or when increasing the number of judges reviewing each paper. This also holds even when we change one of the convex functions used in the Separation-Deviation model. However, when we choose different convex functions to be applied to different papers depending on group membership can help correct for bias, and we can outperform taking the mean and using only quadratic convex functions. We also show a mathematical relationship between taking the mean and the Separation-Deviation model that uses quadratic functions, and why they yield such similar results. While we believe this simulation is a good start for illustrating the effectiveness of the Separation-Deviation model, there are some caveats. All of the true scores were drawn from a normal distribution and were truncated to two decimal points. Because of this, all of these results only hold for this specific situation. Also, the way that we assign bias distributions to author-judge pairings may not be entirely analogous to a real-world situation. In our simulation, we only have two groups where a judge will apply a negative bias against papers that aren’t in their group. This may not be entirely reflective of how bias functions in real life. As demonstrated by Williams et al. [7], a person who is part of one marginalized group faces a different kind of bias than someone who is part of multiple marginalized groups. In our simulation, this might be represented as having more than two groups, where the bias some groups face from other judge groups is next to nothing, and the bias some other groups face gets compounded from other judge groups. Additionally, more work should be done to see how changing the distribution the true scores are drawn from, the number of judge groups, and the number of paper groups, impacts the effectiveness of the Separation-Deviation model compared to finding mean scores.

## 9 Acknowledgements

I would like to express my deepest gratitude to Dr. Swati Gupta for guiding and advising me throughout this endeavor. I would also like to thank Dr. Sahil Singla for providing useful feedback for this project. Lastly, I would like to thank Anubha Mahajan, Jessica Hernandez, and Ramya Raja for providing insightful suggestions at various stages of the project.

## References

- [1] Dorit S. Hochbaum and Asaf Levin. “Methodologies and Algorithms for Group-Rankings Decision”. In: *Manage. Sci.* 52.9 (Sept. 2006), pp. 1394–1408. ISSN: 0025-1909. DOI: 10.1287/mnsc.1060.0540. URL: <https://doi.org/10.1287/mnsc.1060.0540>.
- [2] Dorit S. Hochbaum and Erick Moreno-Centeno. “Joint aggregation of cardinal and ordinal evaluations with an application to a student paper competition”. In: *CoRR* abs/2101.04765 (2021). arXiv: 2101.04765. URL: <https://arxiv.org/abs/2101.04765>.
- [3] Andi Peng et al. “What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring”. In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (Oct. 2019), pp. 125–134. DOI: 10.1609/hcomp.v7i1.5281. URL: <https://ojs.aaai.org/index.php/HCOMP/article/view/5281>.
- [4] Michael L Platt and Scott A Huettel. “Risky business: The neuroeconomics of decision making under uncertainty”. In: *Nature Neuroscience* 11.4 (2008), pp. 398–403. DOI: 10.1038/nn2062.
- [5] Manish Raghavan et al. “Mitigating bias in algorithmic hiring”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Jan. 2020. DOI: 10.1145/3351095.3372828. URL: <https://arxiv.org/abs/1902.03731>.
- [6] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. *Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring*. 2020. DOI: 10.48550/ARXIV.2012.00423. URL: <https://arxiv.org/abs/2012.00423>.
- [7] Joan Williams, Katherine Phillips, and Erika Hall. “Double Jeopardy? Gender Bias Against Women of Color in Science”. In: (Jan. 2014). DOI: 10.13140/2.1.1763.8723.

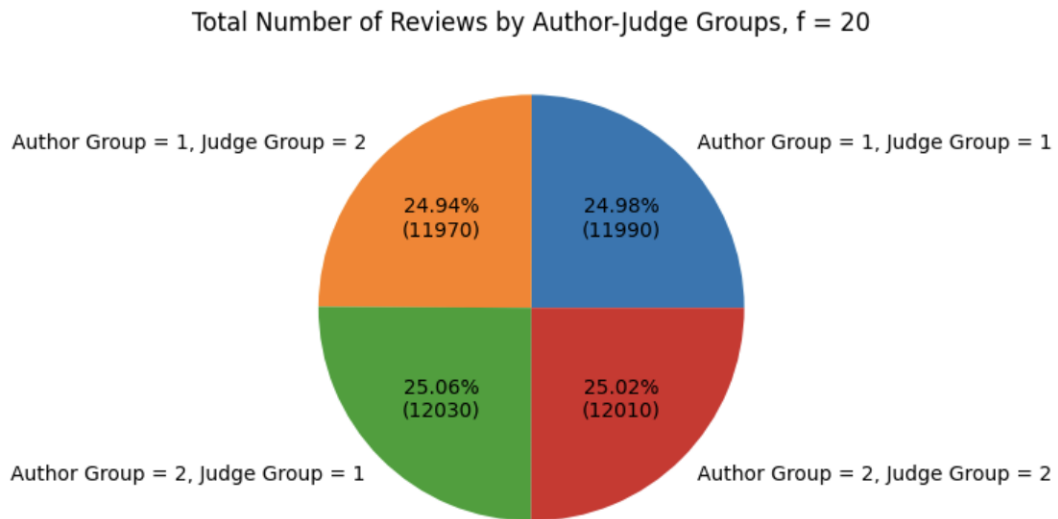
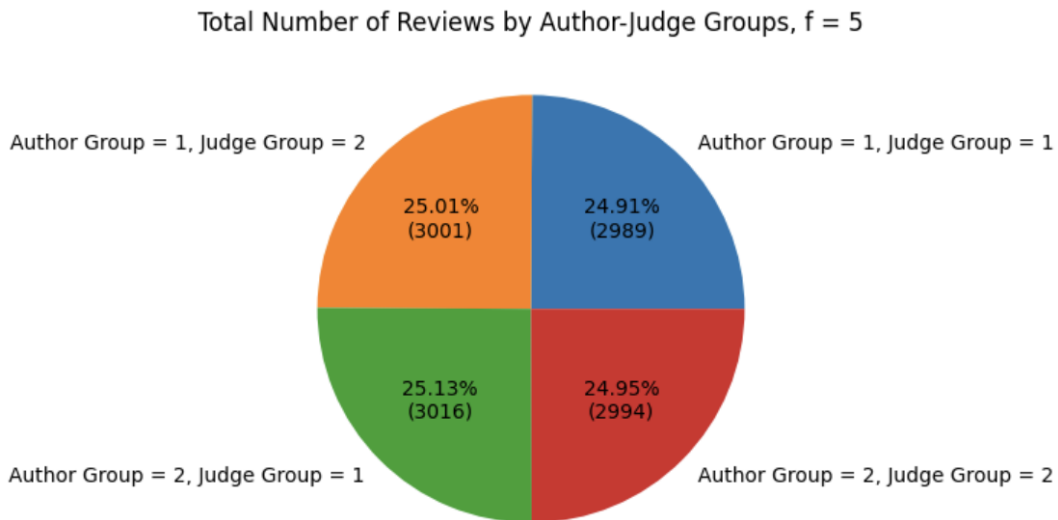


Figure 1: Pie Charts showing the proportions of reviews for each possible author-judge group combination for Scenario 1

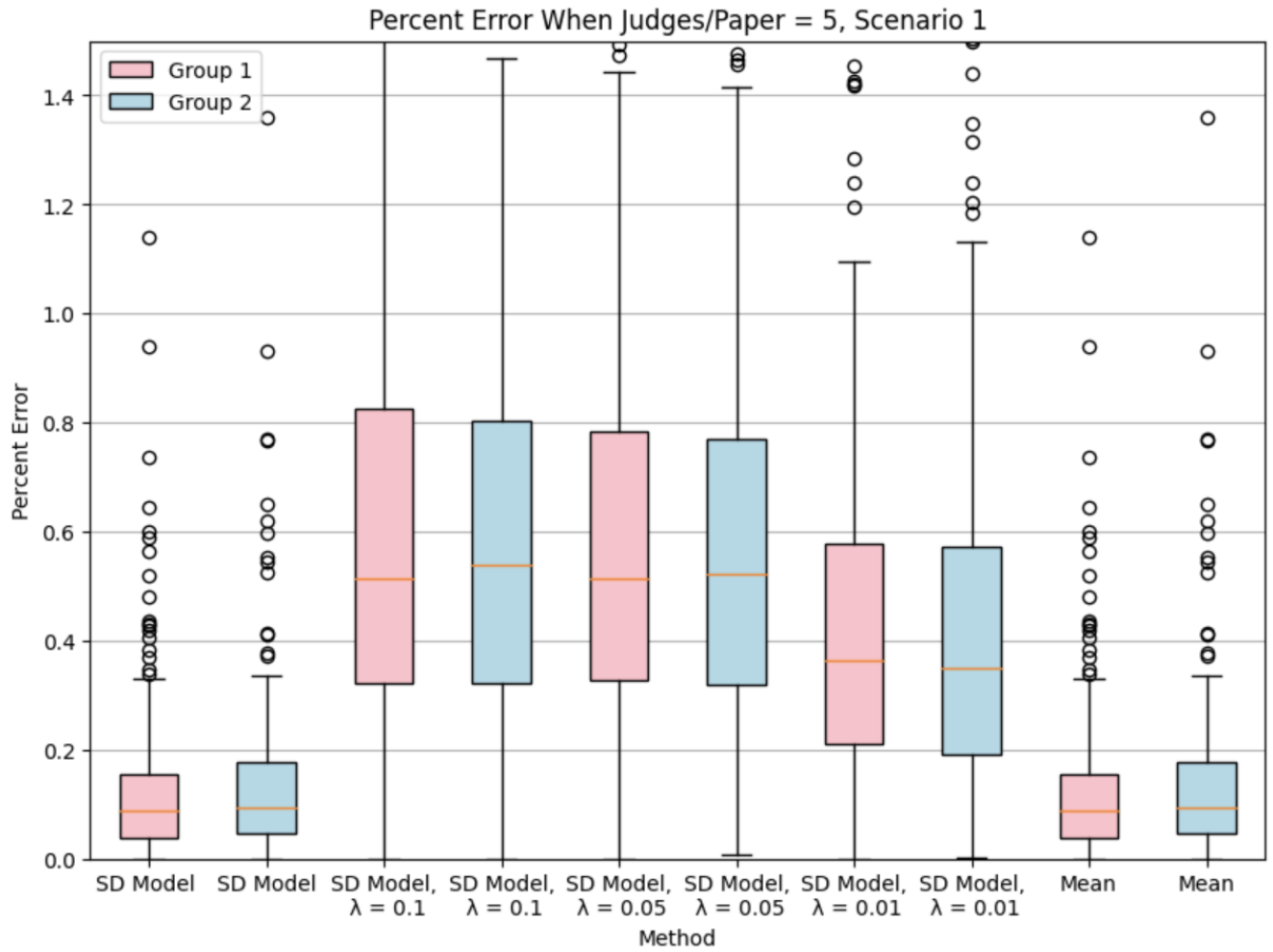


Figure 2: Box plots plotting the percent error across all trials for each group for Scenario 1 where  $f = 5$ .

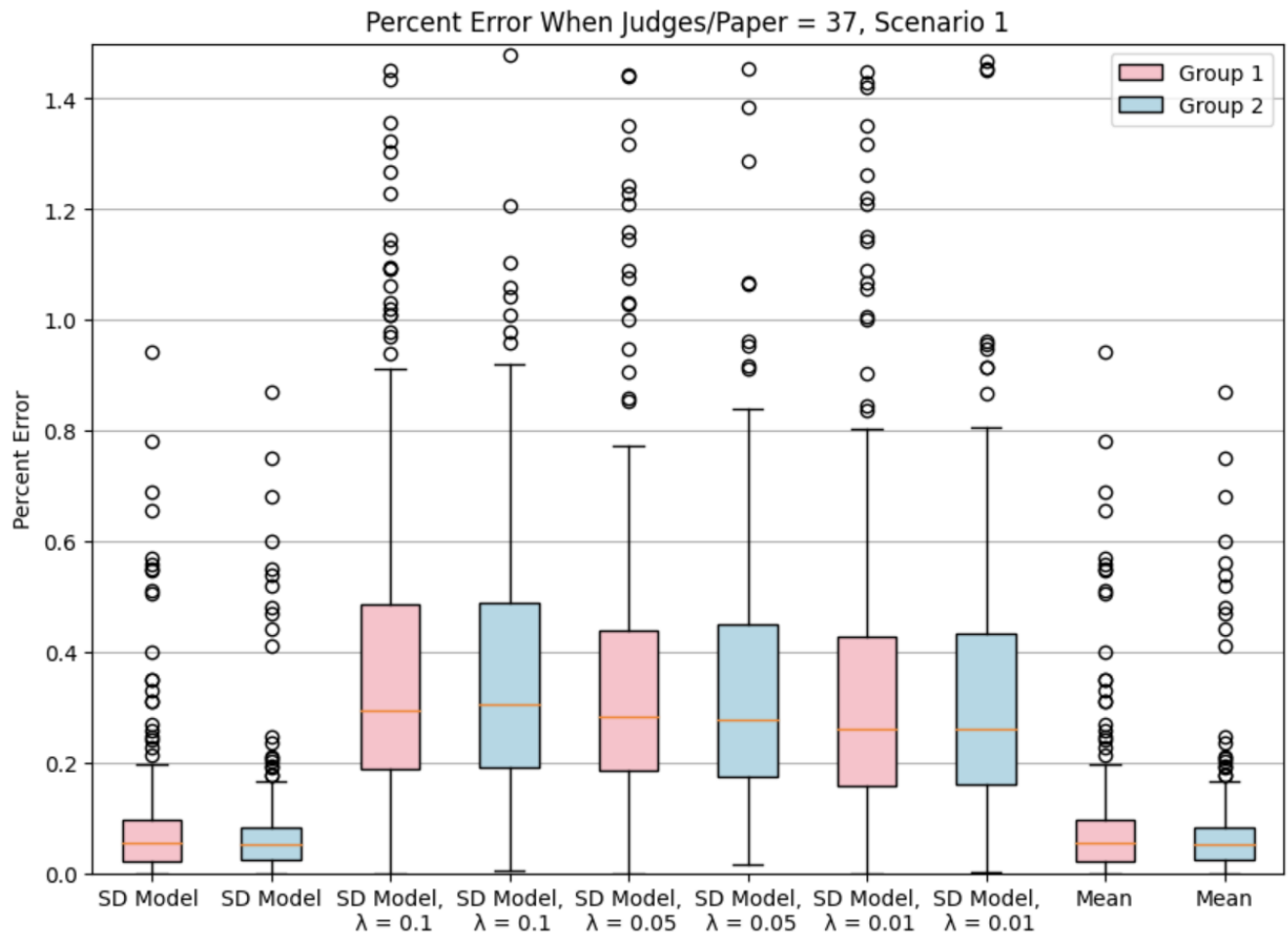


Figure 3: Box plots plotting the percent error across all trials for each group for Scenario 1 where  $f = 20$ .

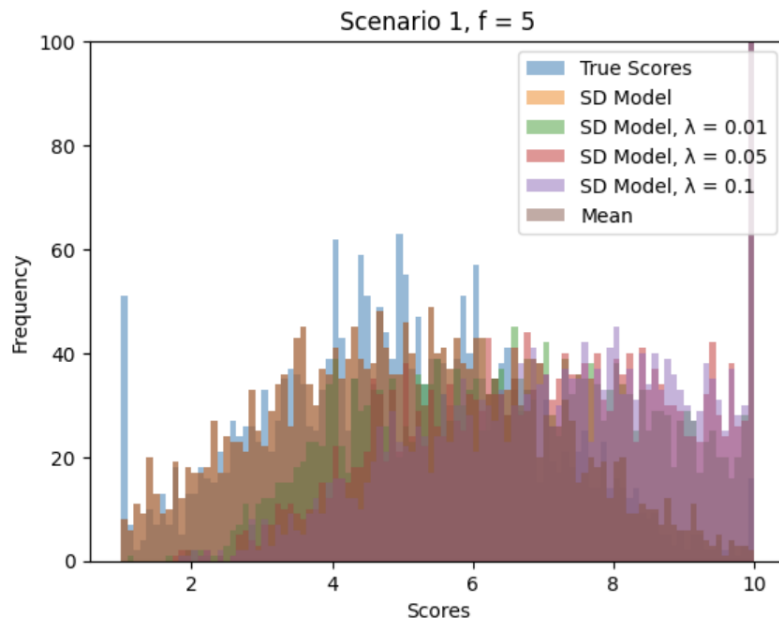


Figure 4: Histogram for Scenario 1 comparing the outputs of the SD model and the mean to the true scores when Judges/Paper = 5

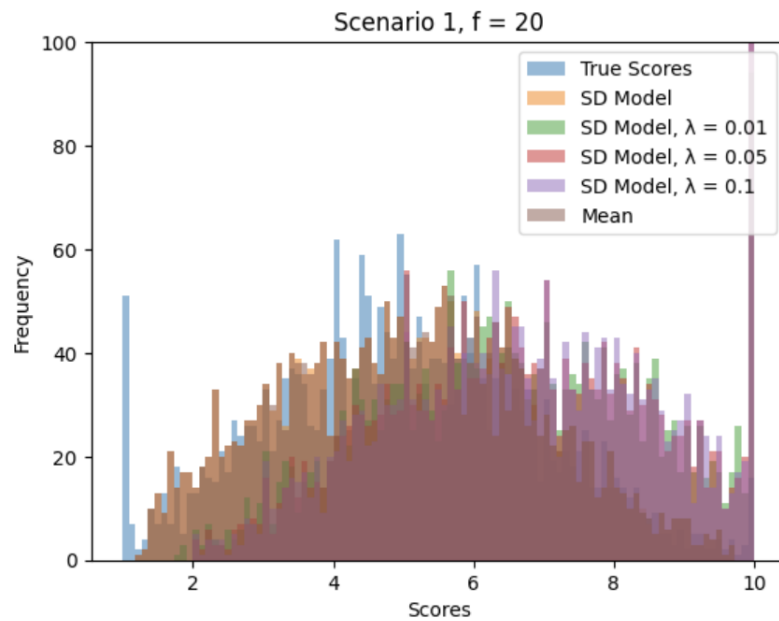


Figure 5: Histogram for Scenario 1 comparing the outputs of the SD model and the mean to the true scores when Judges/Paper = 20

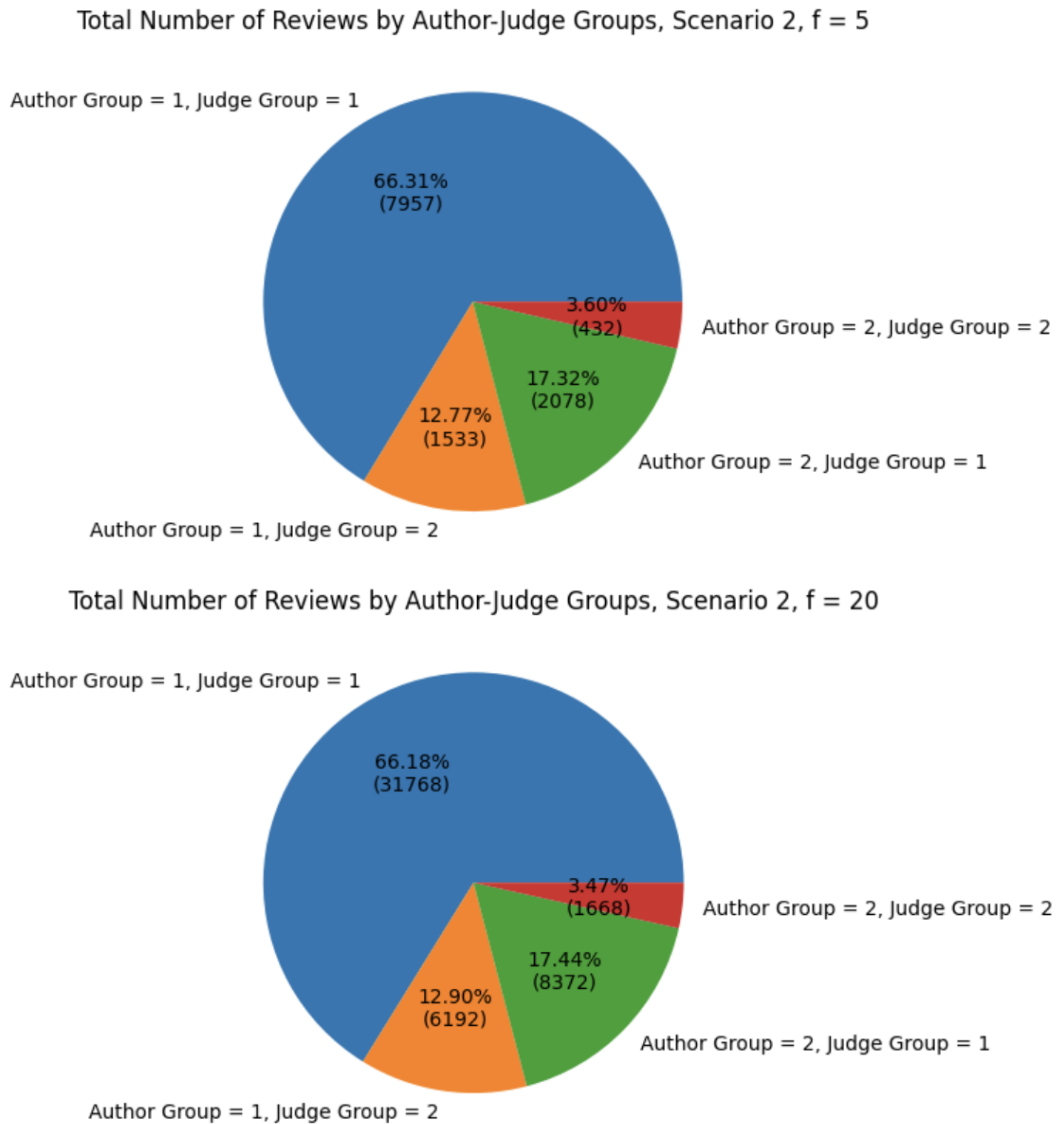


Figure 6: Pie Charts showing the proportions of reviews for each possible author-judge group combination for Scenario 2

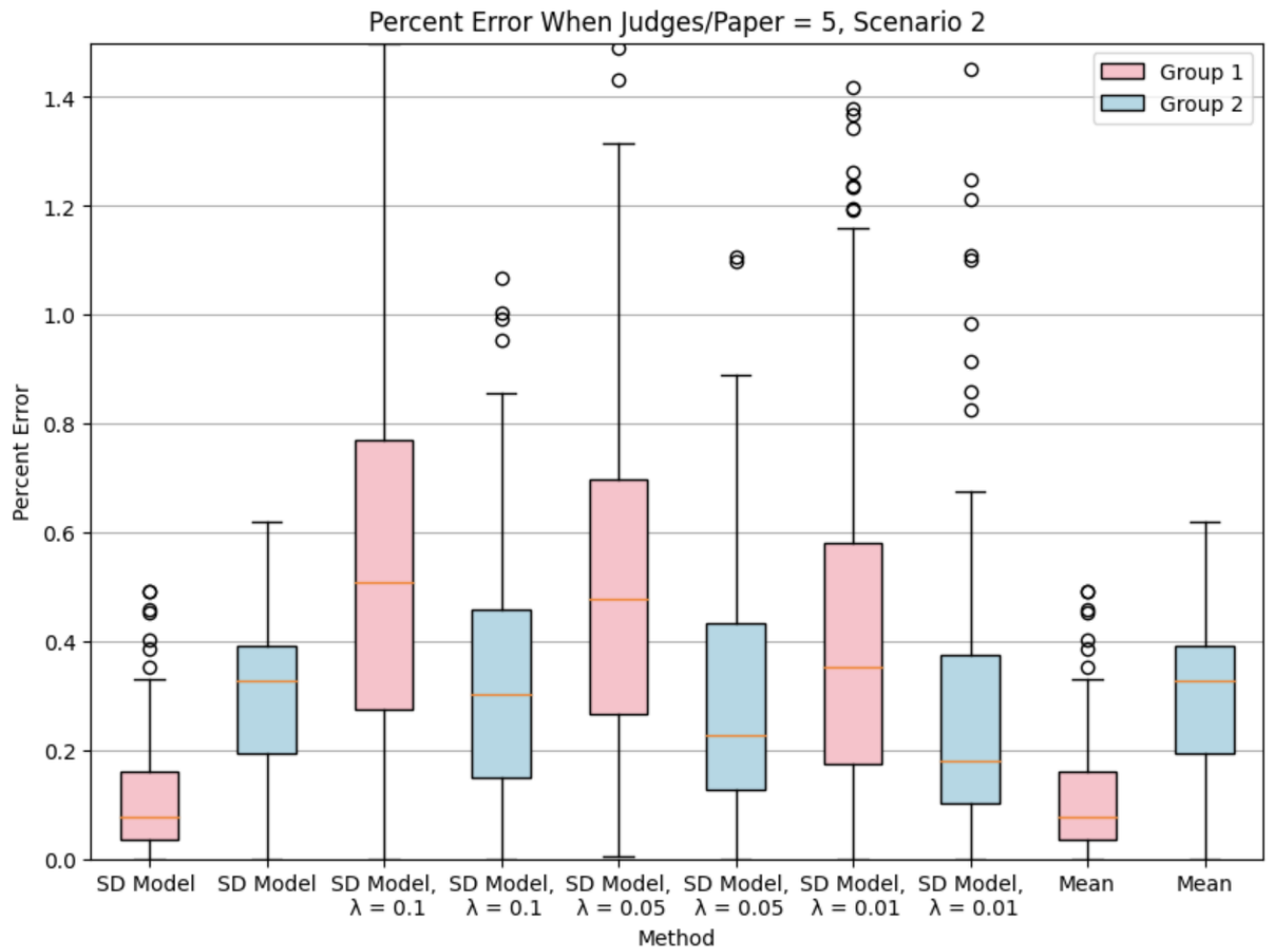


Figure 7: Box plots plotting the percent error across all trials for each group for Scenario 2 where  $f = 5$ .

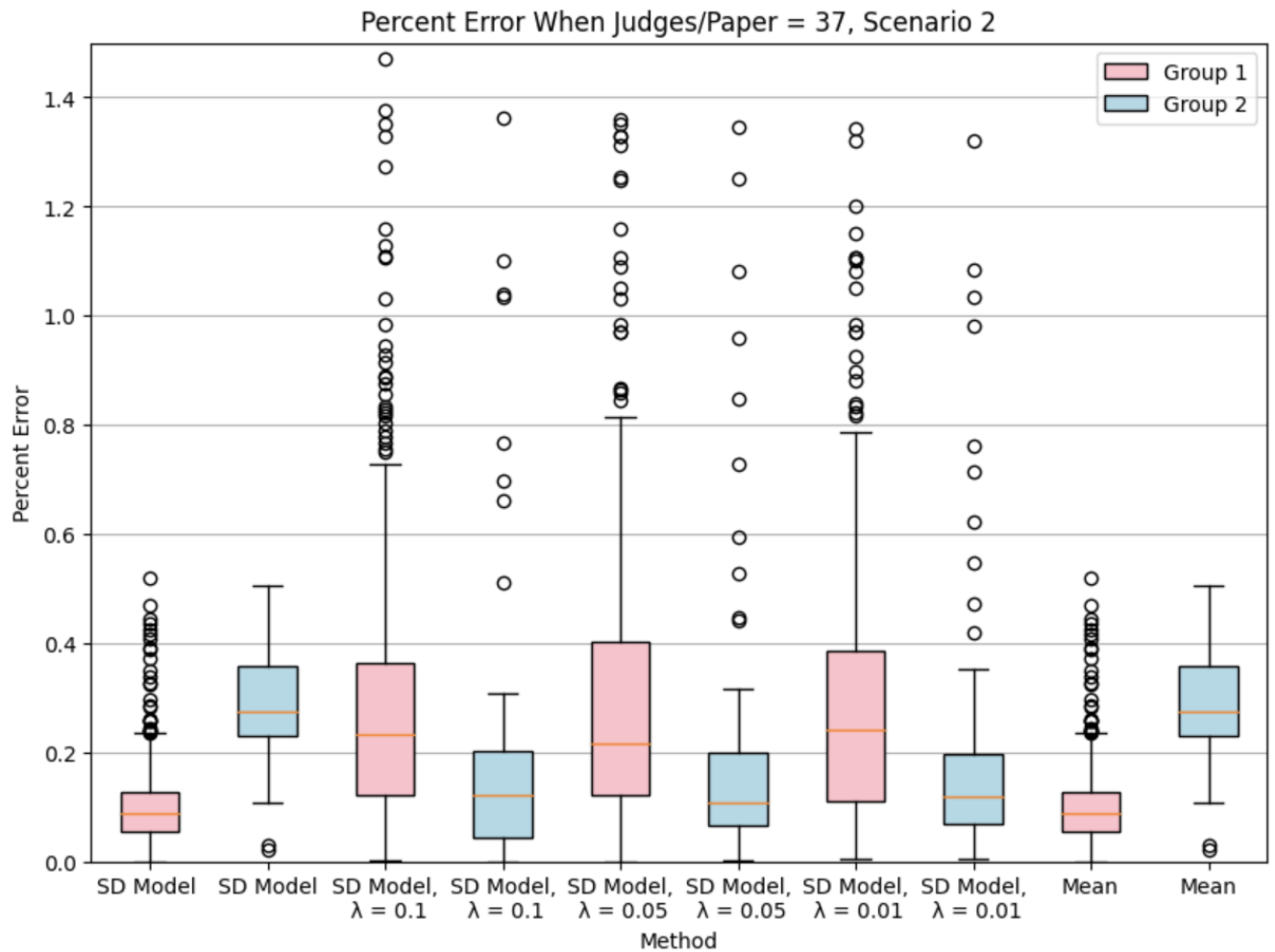


Figure 8: Box plots plotting the percent error across all trials for each group for Scenario 2 where  $f = 20$ .

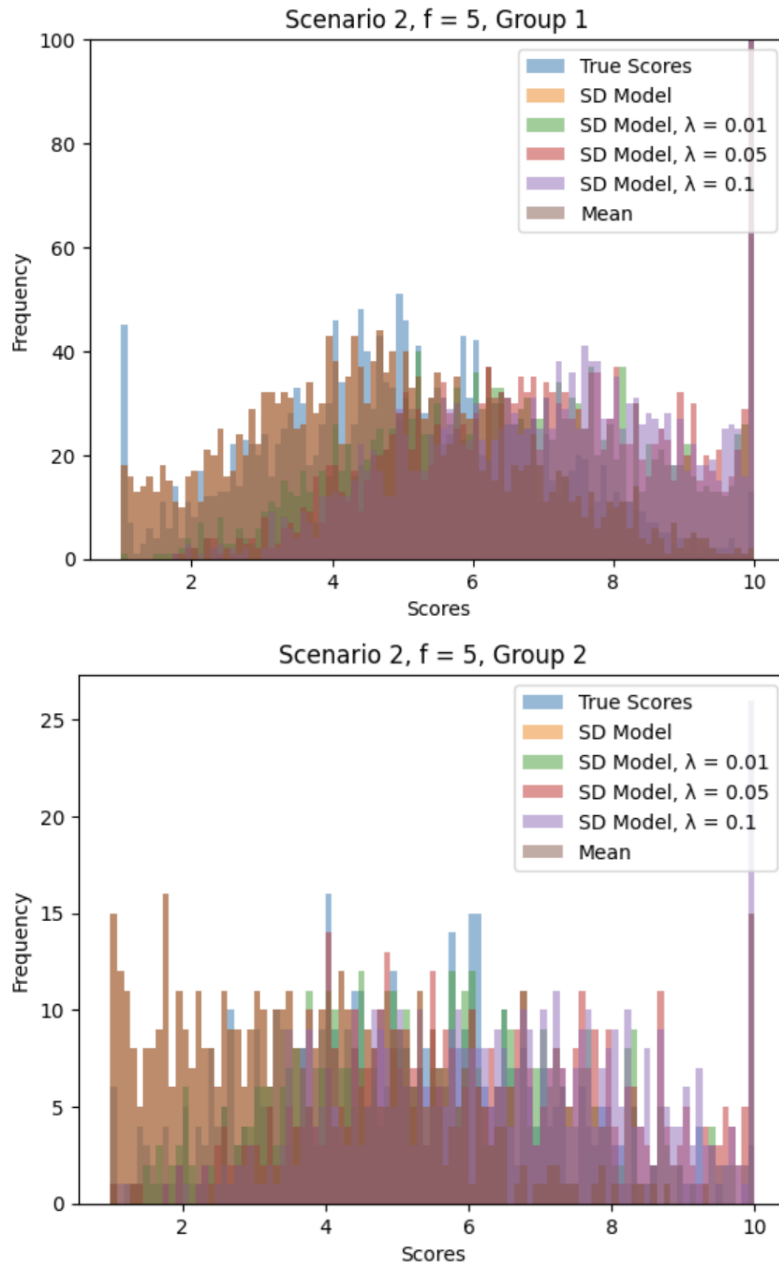


Figure 9: Histograms for Scenario 2 comparing the outputs of the SD model and the mean to the true scores when Judges/Paper = 5

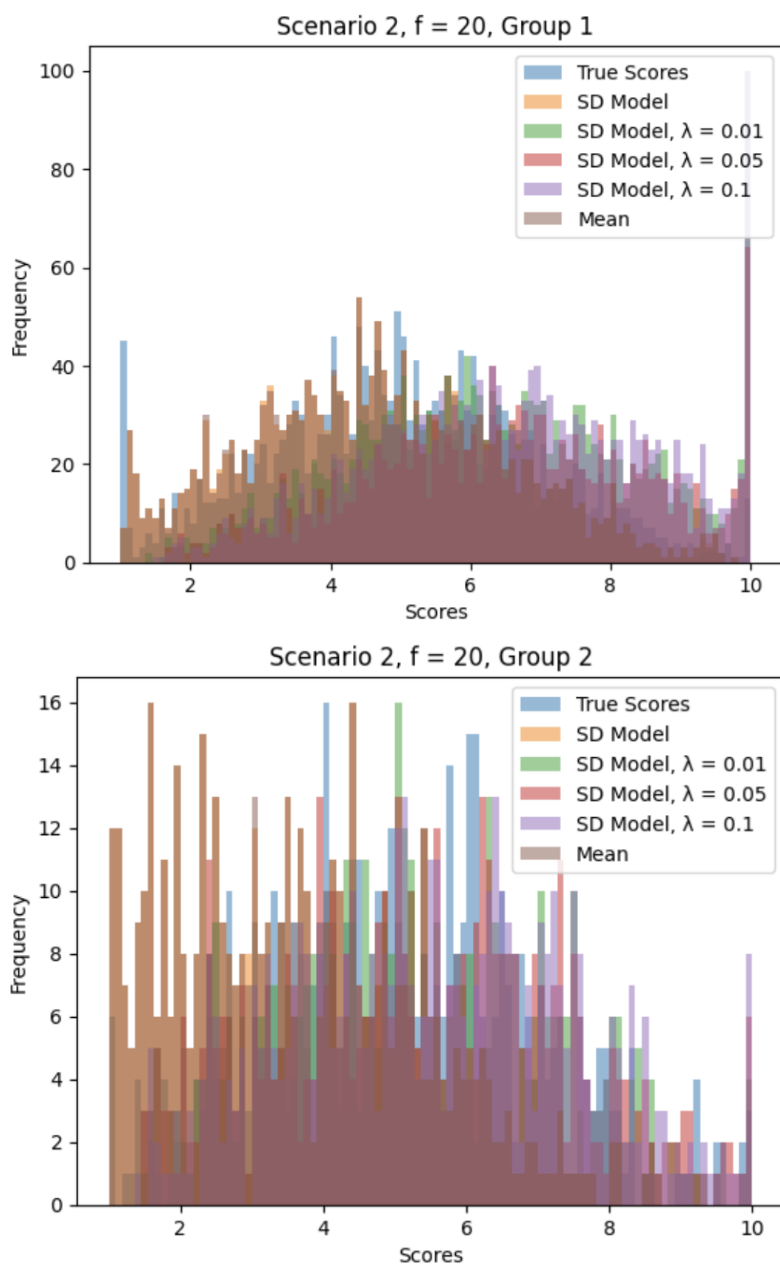


Figure 10: Histograms for Scenario 2 comparing the outputs of the SD model and the mean to the true scores when Judges/Paper = 20

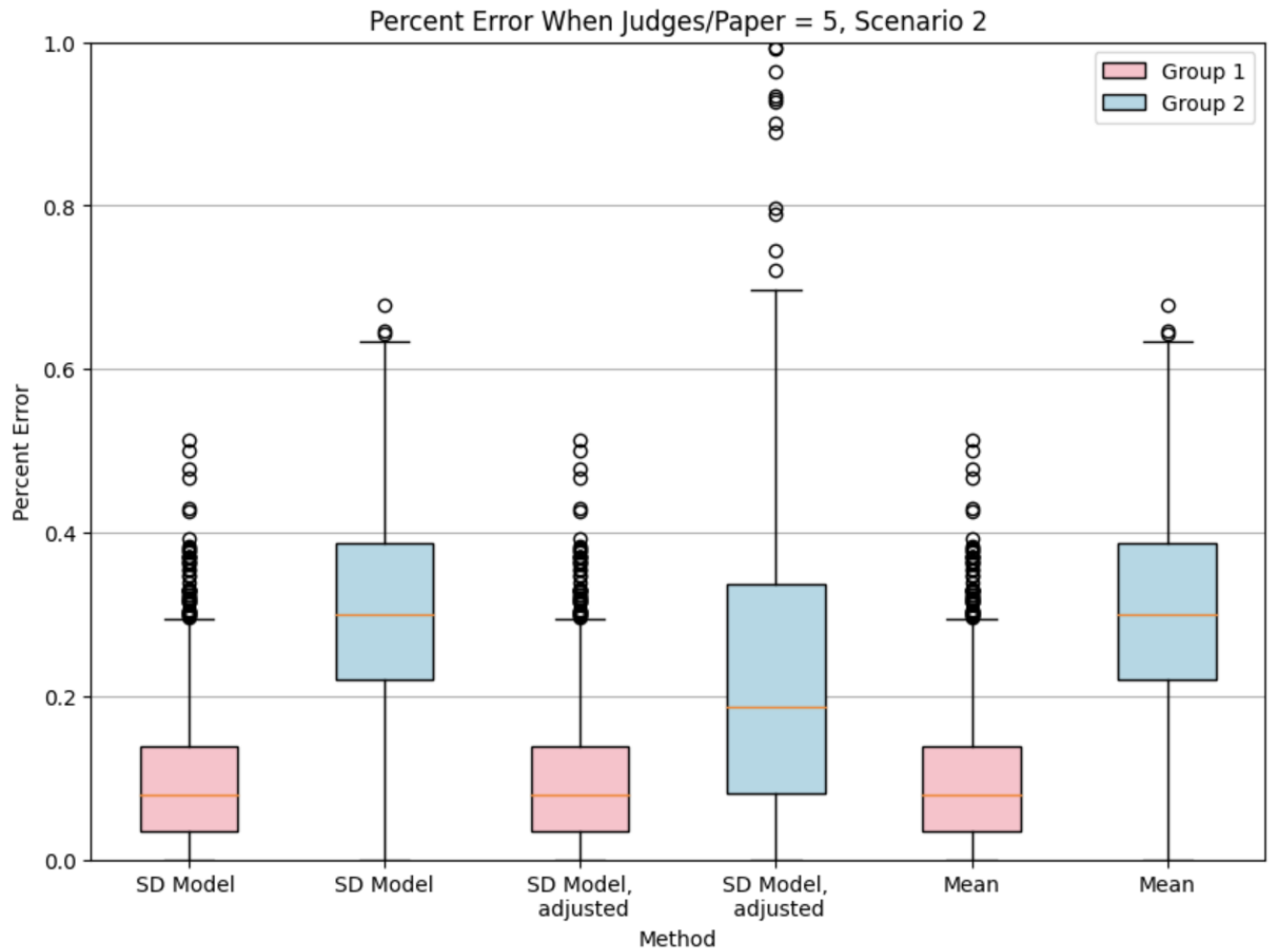


Figure 11: Box plots plotting the percent error across all trials for each group for Scenario 2 where  $f = 5$ , including the adjusted SD model

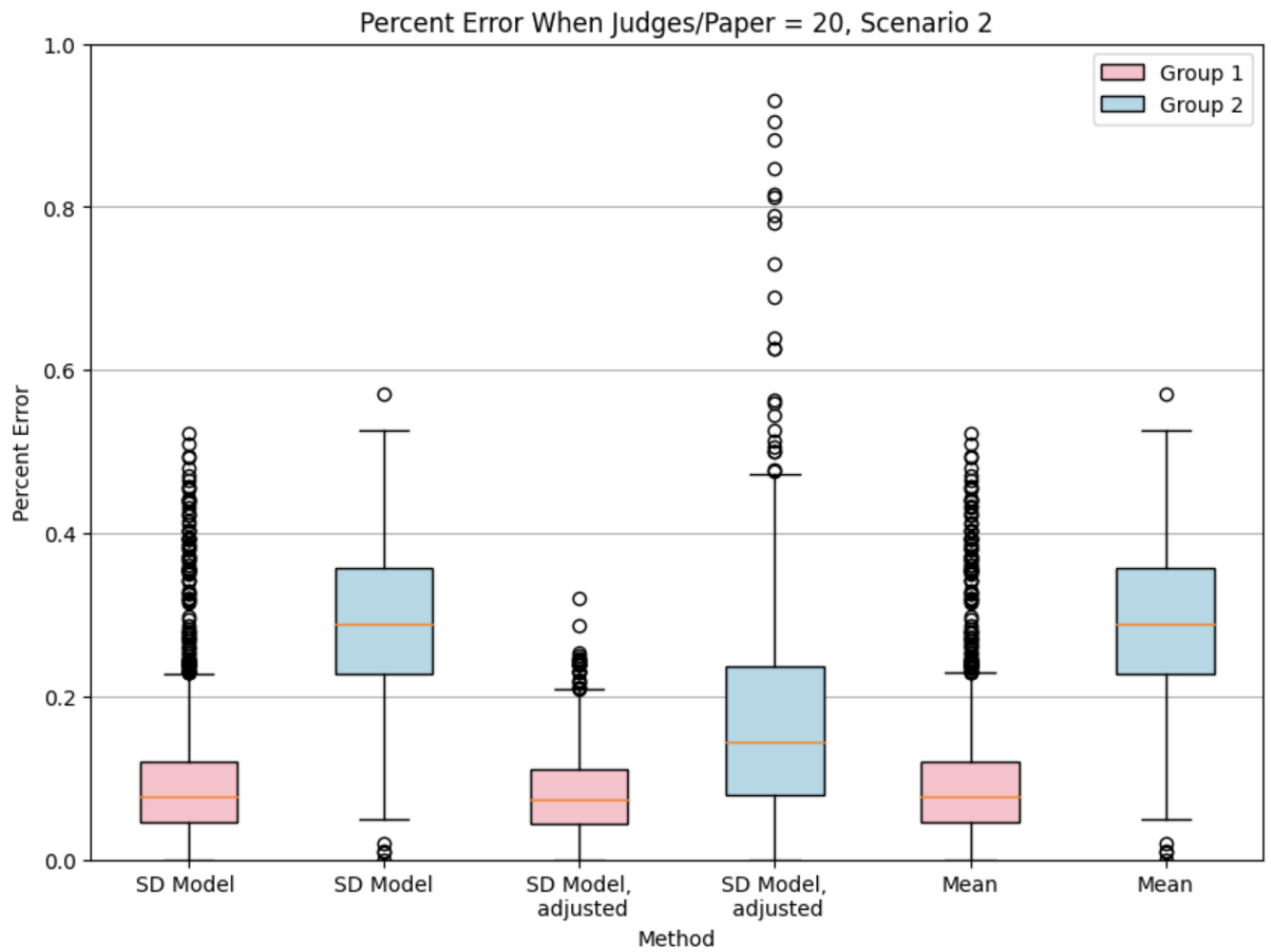


Figure 12: Box plots plotting the percent error across all trials for each group for Scenario 2 where  $f = 20$ , including the adjusted SD model

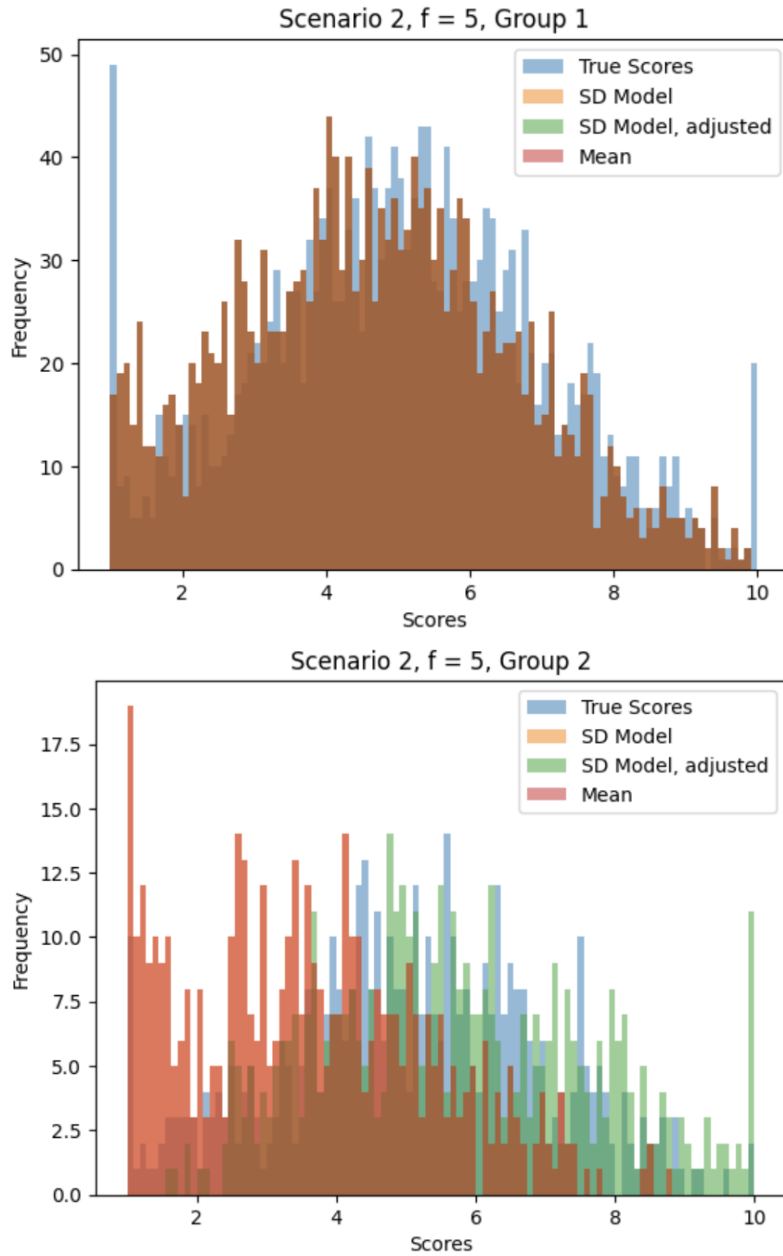


Figure 13: Histograms for Scenario 2 comparing the outputs of the SD model and the mean to the true scores when Judges/Paper = 5, including the adjusted SD model

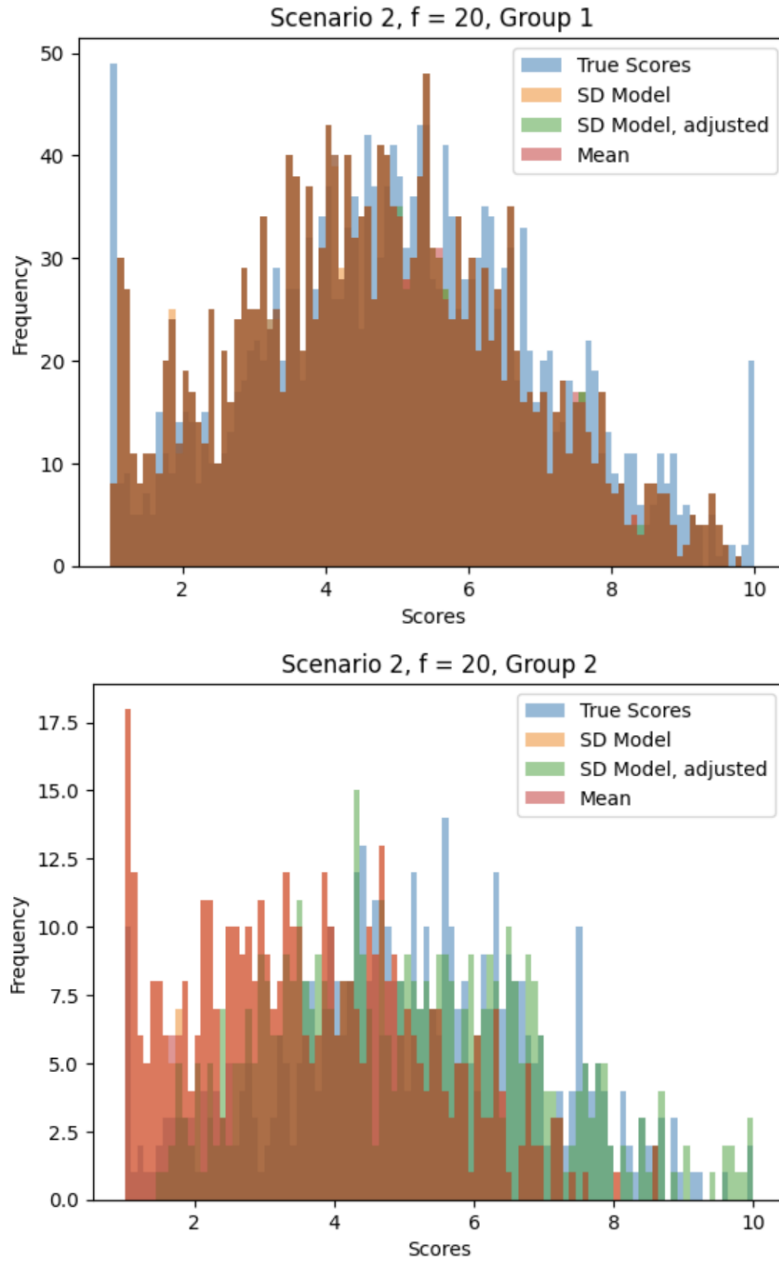


Figure 14: Histograms for Scenario 2 comparing the outputs of the SD model and the mean to the true scores when Judges/Paper = 20, including the adjusted SD model

Paper Index	Average Scores, Scenario 1 f=37	SD Model Scores, Scenario 1 f = 37
0	4.62	4.61
1	5.22	5.22
2	4.66	4.66
3	4.87	4.88
4	4.48	4.47
5	5.25	5.25
6	3.43	3.42
7	4.02	4
8	4.13	4.11
9	6.35	6.33
10	4.29	4.27
11	5.84	5.8
12	3.96	3.93
13	4.96	4.93
14	4.15	4.11
15	3.78	3.74
16	3.07	3.03
17	6.74	6.71
18	5.73	5.68
19	2.76	2.71
20	5.37	5.32
21	3.24	3.2
22	4.55	4.52
23	7.03	7.01
24	4.01	4
25	4.57	4.56
26	7	6.98
27	5.12	5.11
28	5.31	5.29
29	4.29	4.28
30	3.99	3.98
31	4.95	4.94
32	2.68	2.68
33	4.71	4.73
34	5.35	5.39
35	4.49	4.52
36	3.64	3.66
37	5.47	5.5
38	5.03	5.04
39	2.44	2.45
40	4.2	4.2
41	5.6	5.63
42	4.59	4.62
43	5.72	5.76
44	6.86	6.92
45	6.89	6.93
46	5.3	5.34
47	5.62	5.65
48	3.67	3.7
49	4.16	4.2
50	5.94	5.97
51	6.46	6.5
52	3.39	3.43
53	4.6	4.63
54	3.56	3.59
55	4.26	4.29
56	3.51	3.54
57	5.14	5.14
58	4.89	4.89
59	5.98	5.98

Figure 15: Table comparing outputs of one iteration of the SD model when  $g_i^{(k)}$  is quadratic and taking the mean judge score. Highlighted rows are where the two methods yielded different results. Note that these results differ by small quantities (a few hundredths at most).