A Missing Proofs

A.1 Proof of Theorem 3.4

Theorem (Restatement of Theorem 3.4)**.** *A review graph G is recovery-resilient* if and only if *the review graph G has a doubly-connected component S that covers all items, i.e. $C(S) = [N]$.*

Proof. **Proof of Sufficiency.** We first consider the "if" direction. That is, the qualities of all items covered by a doubly-connected component can always be perfectly recovered by any solution of LP (3). Without loss of generality, in the following argument, we fix any given review scores, any solution of LP. Following the design of Algorithm 1, we present a bottom-up induction proof.

Base case: the smallest unit of a doubly-connected component is a single vertex, which represents some reviewer i. According to the linear constraint in LP (3), the recovered qualities x of all items in I_i must be a linear transformation of their review scores y^i by reviewer i. As these review scores y^i are also a linear transformation of the true qualities x^*, so the recovered qualities x solved by the LP (3) must be a linear transformation of the true qualities x^*. In another word, the solution of the LP (3) must perfectly recover the qualities of all items in I_i.

Inductive case: Given two doubly-connected components A, B, and the solution of LP (3) restricted to A, B is a perfect recovery respectively to the items covered by A and B. This means, by definition, we have $k_A, k_B > 0$ and b_A, b_2 such that $x_A = k_A x_A^* + b_A, x_B = k_B x_B^* + b_B$, where x_A, x_A^* and x_B, x_B^* are the recovered *vector* of qualities and true qualities indexed by items in A, B. According to Algorithm 1, if A and B share at least 2 commonly covered items, then their union $D = A \cup B$ is also a doubly-connected component. We show that the solution of (3) must perfectly recover the qualities of items covered by D.

Let u, v be two of the items covered by both component A and B. Denote their recovered (by LP (3) restricted to D) and true qualities be x_u, x_u^* and x_v, x_v^* respectively. Let x_A, x_B denote the recovered x when restricted to A, B respectively. By definition, x_A [resp. x_B] is a feasible solution to LP (3) restricted to A [resp. B]. Our induction hypothesis thus implies that there are the two linear functions $x_A = k_A x_A^* + b_A, x_B = k_B x_B^* + b_B$. Note that x_u, x_v should appear in both vector x_A and x_B. That means these two linear functions $x_A = k_A x_A^* + b_A, x_B = k_B x_B^* + b_B$ intersect at both (x_u, x_u^*) and (x_v, x_v^*). Since item true qualities x^* are assume to be unequal for any two different items, the only possibility when the two linear functions have two intersections is that they are identical, i.e., $k_A = k_B, b_A = b_B$. This implies the entire x_D vector where $D = A \cup B$ must satisfy $x_D = k_A \cdot x_D^* + b_A$. That is, recovered qualities in both x_A, x_B follow the same linear transformation from the true qualities. Therefore, all items covered by D can be perfectly recovered by the solution of LP (3).

Proof of Necessity. For the "only if" direction, we prove its contrapositive statement. That is, if there is no doubly-connected component S that covers all items in the review graph, then there must exist some paper assignment that induces review graph G and some review scores under which the solution to LP (3) cannot perfectly recover the true scores.

We start with a few simplifications, that are without loss of generality. First, according to the above proof of "if" direction, we know that papers within any doubly connected component in the review graph can be perfectly recovered. This means we can replace each of these components by a single vertex, i.e., a single reviewer, with some linear scoring function, who reviewed all the items covered by this doubly connected component. After this transformation, the new review graph will have at most one edge connecting any two vertices. Second, to construct paper assignment that induces this review graph, we will let each edge $e = (i, j)$ in the given review graph correspond to a unique item e, which is only reviewed by reviewer i, j, but no one else.

Let $x \in \mathbb{R}^N$ be an arbitrary vector solution to LP (3) which contains the recovered qualities of all the N items. Without loss of generality, we will pick x as the true quality x^* since we know x^* must be a feasible solution as well. Next we will show there exists another solution $\widetilde{x} \in \mathbb{R}^N$ to the LP (3) such that $\widetilde{x}$ is not linear to x^* (i.e., their corresponding entries do not have linear relation). This implies perfect recoverability is not possible by definition.

To construct $\widetilde{x}$, we will craft a linear function for every reviewer i determined by coefficient k_i and constant b_i. That is, for every item u reviewed by reviewer i, we let $\widetilde{x}_u = k_i x_u^* + b_i$, where k_i, b_i are to be determined later. We wish to set item u's constructed score as $\widetilde{x}_u$. If we could succeed in doing

so, then as long as $k_i \neq k_j$ for all i, j, then $\widetilde{x}$ cannot have a linear relation with x^*, which completes our proof. However, there are some constraints to be satisfied when setting $\widetilde{x}_u$ as item u's constructed score, which is why we have to pick $\{k_i, b_i\}_{i\in[M]}$ and x^* carefully. The constraints come from edges of the graph: each edge $e = (i, j)$ connecting reviewer i and j corresponds to an item e that reviewer i, j both reviewed. This will require the constructed scores, when viewed from i's and j's perspective, have to be consistent, i.e.,

$$k_i x_e^* + b_i = k_j x_e^* + b_j, \quad \forall e$$

which both equal the constructed $\widetilde{x}_e$.

Since recovery-resilience of a review graph needs to hold for any given underlying true qualities, to disapprove it we only need to identify one set of true item qualities to satisfy our construction. Towards that end, we will use the following construction: $k_i = \sqrt{p_i}, b_i = p_i, \forall i \in [M]$, where p_i is the ith smallest prime number from 2. Given these $\{k_i, b_i\}_{i\in[M]}$, we then let

$$x_e^* = -\frac{b_i - b_j}{k_i - k_j} = -\frac{p_i - p_j}{\sqrt{p_i} - \sqrt{p_j}} = -\sqrt{p_i} - \sqrt{p_j}, \qquad \text{for all edge } e = (i, j).$$

It is easy to verify that the above construction does satisfy $k_i x_e^* + b_i = k_j x_e^* + b_j$ for any edge $e = (i, j)$. Moreover, no two edges with $e = (i, j), e' = (i', j')$ can have their quality $\widetilde{x}_e = \widetilde{x}_{e'}$ because for any four $p_i, p_j, p_{i'}, p_{j'}$ with at least two unique prime $p_i, p_{i'}$, it is impossible that $\sqrt{p_i} + \sqrt{p_j} = \sqrt{p_{i'}} + \sqrt{p_{j'}}$. To see this, if we take the square of both side, this will lead to $2\sqrt{p_i p_j} - 2\sqrt{p_{i'} p_{j'}} = p_i + p_j + p_{i'} + p_{j'}$. Now if we take the square of both sides again, we have $-8\sqrt{p_i p_j p_{i'} p_{j'}} = (p_i + p_j + p_{i'} + p_{j'})^2 - 4(p_i p_j + p_{i'} p_{j'})$. However, the RHS is rational, yet the LHS must be irrational since at least $p_i, p_{i'}$ are unique prime number, a contradiction. Therefore, our construction of $\widetilde{x}$ is indeed valid. This concludes that the review graph with no more than one edge cannot be recovery-resilient.

□

A.2 Proof of Theorem 3.6

Theorem (Restatement)**.** *Convex Program (4) is equivalent to LSC with $\mathcal{H} = \mathcal{H}_L(C)$ in the following sense: for any optimal solution $(\mathbf{x}^*, \epsilon^*)$ to (4), there exists $\mathbf{f}^* = \{f_j^* \in \mathcal{H}_L(C)\}_{j\in[M]}$ such that $(\mathbf{x}^*, \epsilon^*, \mathbf{f}^*)$ is optimal to (2).*

Proof. For any optimal solution $(\mathbf{x}^*, \epsilon^*)$ to (4), we can linearly interpolate points $\{(x^*(I_j^\ell)+\epsilon_j^{\ell*}, y_j^\ell)\}_l$ and construct the linear function $f_j^*(x) = \alpha x + \beta$ with $\alpha = \frac{y_j^\ell - y_j^{\ell-1}}{\widetilde{x}_j^\ell - \widetilde{x}_j^{\ell-1}}$ and $\beta = y_j^\ell - \alpha x^*(I_j^\ell)$, for all $j \in [M], 2 \leq \ell \leq |I_j|$. Therefore, $(\mathbf{x}^*, \epsilon^*, \mathbf{f}^* = \{f_j^*\}_{j\in[M]})$ is a *feasible solution* to LSC and thus the optimal objective of LSC is at least that of LP (4).

To show that the optimal objective of LCS is at most that of LP (4), observe that any feasible solution to LSC must satisfies the *linear equality constraint* of LP (4) because

$$\frac{1}{\alpha_j} = \frac{\widetilde{x}_j^\ell - \widetilde{x}_j^{\ell-1}}{y_j^\ell - y_j^{\ell-1}} = \frac{\widetilde{x}_j^{\ell+1} - \widetilde{x}_j^\ell}{y_j^{\ell+1} - y_j^\ell} \quad \forall j \in [M], 2 \leq \ell \leq |I_j|$$

which is precisely the *linear equaltiy constraints* of LP (4). This implies that the feasible region of LP (4) contains the feasible region of LSC restricted to $\mathbf{x}, \epsilon$. Therefore, the optimal objective of LSC is also at most that of LP (4), as desired. □

A.3 Proof of Theorem 4.2

Theorem (Restatement)**.** *The ℓ_2 Matrix Seriation (4.1) problem can be solved by the following Functional Optimization Problem.*

$$\begin{aligned} \min_{\mathbf{x},\mathbf{f}} \quad & \sum_{j=1}^{M} \sum_{\ell=1}^{|I_j|} (\epsilon_j^\ell)^2 && (8) \\ \text{s.t.} \quad & y_j^\ell = f_j(x(I_j^\ell)) + \epsilon_j^\ell && \forall j \in [M], \ell \leq |I_j| \\ & f_j \in \mathcal{H}_{mono} && j \in [M] \end{aligned}$$

Proof. We can represent reviewers' scores as a matrix $A \in \mathbb{R}^{m \times n}$ where each row represents the scores given by a specific reviewer, and each column represents the scores received by a specific item. Therefore, $A_{i,j}$ represents reviewer i's score for item j. Note that $A_{i,j}$ is a partial matrix. $A_{i,j}$ is empty if reviewer i does not review item j. Starting with an empty matrix B, we fill B_{j,I_j^ℓ} with $f_j(x(I_j^\ell)) = y_j^\ell - \epsilon_j^\ell$ for all $j \in [M], \ell \leq |I_j|$, which is obtained from the solution of the Functional Optimization Problem (8).

Note that

$$\|A - B\|_2 = \sum_{j=1}^{M} \sum_{\ell=1}^{|I_j|} (y_j^\ell - f_j(x(I_j^\ell)))^2 = \sum_{j=1}^{M} \sum_{\ell \leq |I_j|} (\epsilon_j^\ell)^2$$

which is exactly what FOP (8) minimizes.

In addition, because all scoring functions are monotonically increasing, it means each row of B preserves the order according to items' qualities x. In other words, for any $i, j \in [m], p, q \in [n]$ we have

$$B_{i,q} \leq B_{i,p} \iff B_{j,q} \leq B_{j,p}$$

Therefore, (8) solves the matrix seriation problem. □

B Additional Details and Results for Experiments

In this section, we provide a detailed description of our data generation procedure, as well as various experiments to better understand the strength and limitation of our calibration framework.

B.1 Dataset Generation

Our synthesized data generation follows the procedure below. For the ease of reproduction, we also include the implementation details in the our released code in supplemental materials.

Paper qualities. Without loss of generality, we restrict the true qualities of papers to be within $[0, 10]$. We randomly sample the true quality of each paper from a Gaussian distribution $x \sim \mathcal{N}(5, 1.6)$. This choice of parameter ensures roughly 99.8% of the true qualities fall within $[0.2, 9.8]$. We truncate the value to 0 or 10 if any sample falls below or above the $[0, 10]$ interval.

Paper assignment and review graph. We randomly assign each paper to a pool of reviewers under the natural constraint that a paper should be reviewed at least $\lfloor kM/N \rfloor$, according to the Algorithm 2. Meanwhile, for the experiment shown in Figure 2, we generate the review graph of double connectivity, according to Algorithm 3. The function $\texttt{choose}(S, k)$ there is to randomly sample from the set S for k elements without replacement.

Paper scores. We randomly assign a scoring function to each reviewer to compute their review scores.

To generate a linear scoring function, $f(x) = kx + b$, we draw the parameter $k \sim \mathcal{U}(0, 2)$ and $b \sim \mathcal{N}(0, 2)$.

To generate a concave function, we take a random linear combination between a set of monotone concave functions $\{c_2 \cdot \sqrt{x}, c_3 \cdot \sqrt[3]{x}, c_4 \cdot \sqrt[4]{x}\}$, where the weight of each function is sampled uniformly random according to $c_p \sim \mathcal{U}(1, \frac{10}{\sqrt[p]{10}}), \forall p \in \{1, 2, 3, 4\}$.

To generate a convex function, we take a random linear combination between a set of monotone convex functions $\{c_1 \cdot x^2, c_2 \cdot x^{2.5}, c_3 \cdot x^3\}$ where the weight of each function is sampled uniformly random according to $c_1, c_2, c_3 \sim \mathcal{U}(0, 1)$.

To generate an arbitrary monotone function, we randomly sample k values from $[0, 10]$ in increasing order and assign them to papers in the corresponding order.

In addition, for the noisy case, a zero-mean Gaussian error $\epsilon \sim \mathcal{N}(0, \sigma)$ is added to the true qualities of the papers for each reviewer-paper pair as the percetion error before applying each reviewer's scoring function.

Algorithm 2 `Random Assignment`

input N papers, M reviewer, k papers per reviewer

$b \leftarrow \lfloor kM/N \rfloor$ # minimum number of reviews requires for each paper
$V \leftarrow \emptyset$ # set of papers that have met the minimum requirement
$\texttt{usage} \leftarrow \{\}$ # map for the number of reviews of each paper
for $i \in \{1 \dots M\}$ **do**
  $S_1 \leftarrow \texttt{choose}\,([N]/V, \min(k, N - |V|))$
    # sample as many papers that have not met the minimum requirement
  $S_2 \leftarrow \texttt{choose}\,(V, k - |S_1|))$
    # sample the remaining papers so the reviewer gets k papers
  $T_i \leftarrow S_1 \cup S_2$
  for $j \in I_i$ **do**
    $\texttt{usage}[j] \leftarrow \texttt{usage}[j] + 1$
    if $\texttt{usage}[j] >= b$ **then**
      $V \leftarrow V + j$
    end if
  end for
end for
return $\{I_i\}_{i \in [M]}$

Algorithm 3 `Random Assignment with Double Connectivity`

input N papers, M reviewer, $k > 2$ papers per reviewer

$b \leftarrow \lfloor kM/N \rfloor$ # minimum number of reviews requires for each paper
$\texttt{usage} \leftarrow \{\}$ # map for the number of reviews of each paper
$V \leftarrow \emptyset$ # set of papers that have met the minimum requirement
$I_1 \leftarrow \texttt{choose}([N], k)$ # sample any k papers for reviewer 1
$T \leftarrow I_1$ # set of papers that have been assigned
$\texttt{usage}[j] \leftarrow 1, \quad \forall j \in I_1$
for $i \in \{2 \dots M\}$ **do**
  $S_1 \leftarrow \texttt{choose}([N]/T, \max(k - 2, N - |T|))$
    # sample as many unassigned papers
  $S_2 \leftarrow \texttt{choose}(T/V, \min(2, |T| - |V|))$
    # sample as many as 2 assigned papers that have not met minimum requirement
  $S_3 \leftarrow \texttt{choose}(V, k - |S_1| - |S_2|)$
    # sample the remaining papers so the reviewer gets k papers
  $I_i \leftarrow S_1 \cup S_2 \cup S_3$
  $T \leftarrow T \cup I_i$
  for $j \in I_i$ **do**
    $\texttt{usage}[j] \leftarrow \texttt{usage}[j] + 1$
    if $\texttt{usage}[j] >= b$ **then**
      $V \leftarrow V + j$
    end if
  end for
end for
return $\{I_i\}_{i \in [M]}$

B.2 Experiments for Remark 3.5

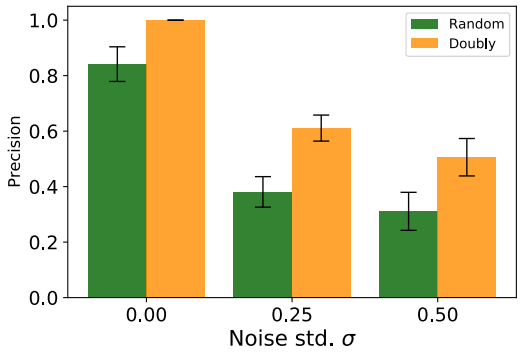

Figure 2: Performance comparisons of LSC with linear constraints between **randomly generated review graph** (green bar) and **randomly generated doubly connected reviewed graph** (orange bar) under different level of noise scale with $N = 1000, M = 350, k = 5$.

In Figure 2, we are able to directly verify our Theorem 3.4 that a randomly generated graph of double connectivity is indeed recovery-resilient, as it is able to achieve the bona fide perfect precision in the noiseless setting. In addition, the topological structure of double connectivity is also less prone to the perception noise, compared to a review graph generated from a completely random assignment.

B.3 Additional Experiments and Discussions on the Effects of Prior Knowledge

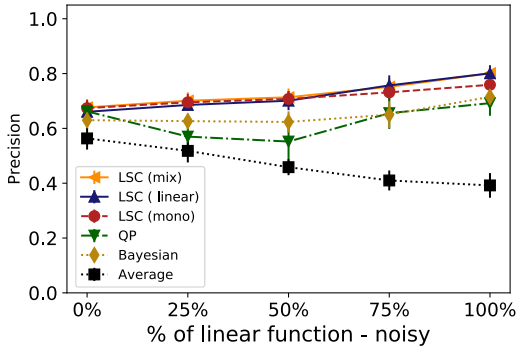

Figure 3: Performance comparisons in the noisy ($\sigma = 0.5$) and mixed setups with linear scoring functions and arbitrary monotone functions. Only LSC (mix) has prior knowledge of every reviewer's scoring function type.

In Section 5.3, we demonstrate the robustness of our model under the situation when the prior knowledge is misspecified. In Figure 3, we can see that, under the noisy setting, our model still outperforms all the baselines. However, it turns out that the influence of whether the prior knowledge is misspecified or not is rather minimal, compared the noiseless setting. Similar pattern is observed in Table 3, where the scoring functions consist of $1/3$ monotonic increasing function, $1/3$ convex functions, and $1/3$ concave functions (all randomly generated). Only LSC (mix) has the exact prior knowledge of every reviewer's scoring function type. The LSC model can effectively utilize such the prior knowledge. We can see significant improvement of the precision under noiseless setting, though such improvement is minor in the more noisy setting. On one hand, our model is indeed able to fit for a wide range of function classes so that more prior knowledge can lead to better calibration. On the other hand, these observations also suggest that with the presence of more perception noise, it could be more beneficial to pick a robust model rather than fixating on an accurate prior knowledge for better calibration results. This justifies the usage of our model for the cases even without strong prior knowledge.

Model \ Metric	Pre. (%)	Avg. Gap	Pre. (%)	Avg. Gap
Average	39.9 ± 2.5	0.70 ± 0.06	38.6 ± 3.5	0.76 ± 0.07
QP	76.4 ± 3.0	0.13 ± 0.02	69.8 ± 4.5	0.21 ± 0.05
Bayesian	54.4 ± 3.1	0.40 ± 0.03	51.1 ± 3.6	0.48 ± 0.04
LSC (mono)	76.4 ± 3.9	0.13 ± 0.03	68.8 ± 3.2	0.22 ± 0.02
LSC (linear)	76.4 ± 3.8	0.14 ± 0.02	67.7 ± 3.1	0.22 ± 0.03
LSC (mixed)	$\mathbf{93.2 \pm 2.5}$	$\mathbf{0.02 \pm 0.01}$	$\mathbf{71.1 \pm 3.4}$	$\mathbf{0.18 \pm 0.03}$

Table 3: Experimental results for **mixed scoring functions** setting on the baselines and LSC with monotone, linear or mixed (given the prior knowledge) constraints. Each entry contains the mean and standard deviation over 20 trials. The table on the left and right side respectively shows the results in the **noiseless** ($\sigma = 0$) and **noisy** ($\sigma = 0.5$) setting.

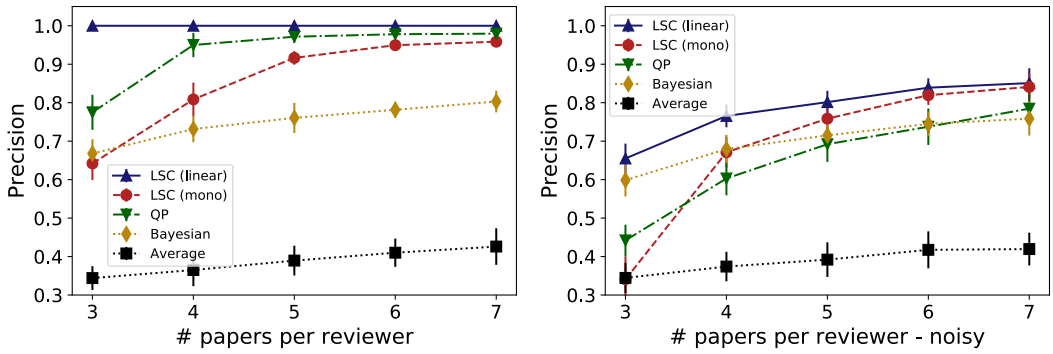

Figure 4: The experiments evaluate the performance of baseline models as well as the LSC under monotone or linear constraints. The first plot shows how the algorithms' performance changes according to k, **the number of papers assigned to each reviewer** in the **noiseless** ($\sigma = 0$) setting, the second plot shows the performance change in the **noisy** ($\sigma = 0.5$) setting.

B.4 Additional Empirical Study

In real applications such academic conferences, the calibration framework could face very different setups. For example, the PKC-2020 [14] conference has 180 submission with an acceptance rate of 0.24 while the AAAI-2020 [2] conference has 8800 submitted papers. It is crucial to investigate and understand the performance of our model at different scales of hardness. Therefore, we conduct a set of empirical studies with datasets generated by different hyperparameters.

The number of papers k assigned to each reviewer Since reviewers can professionally review a large amount of papers in practice, we compare the algorithm performance under different k, changing from 3 to 7, as shown in the first column of Figure 4. We observe improved performance as the number k of papers per reviewer increases. Interestingly, it turns out that $k = 6$ serves as a sweet spot to balance reviewers' workload and the calibration performance since after $k = 6$ the calibration quality starts to increases only mildly. Our model LSC (linear) can always perfectly recover the true qualities of the best papers under the noiseless case even when each reviewer only reviews 3 papers.

The paper to reviewer ratio ($N : M$) While we assume a $1 : 1$ ratio between the number of papers and reviewers in our standard setup, there is sometimes less reviewers than papers in large academic conferences.[9] To understand the algorithm performance under different paper to reviewer ratios, we test a $3 : 2$ ratio and $2 : 1$ ratio by fixing other hyperparameters (e.g., $k = 5, N = 1000$). As can be seen in Figure 5, comparing to the performance drop with a smaller k in the Figure 4, the decrease in the number of reviewer has similar effect on our models, but is a relatively less important

[9] For example, NeurIPS-2019 [25] receives 6743 submissions and has around 4500 reviewers, leading to a paper to reviewer ratio around $3 : 2$.

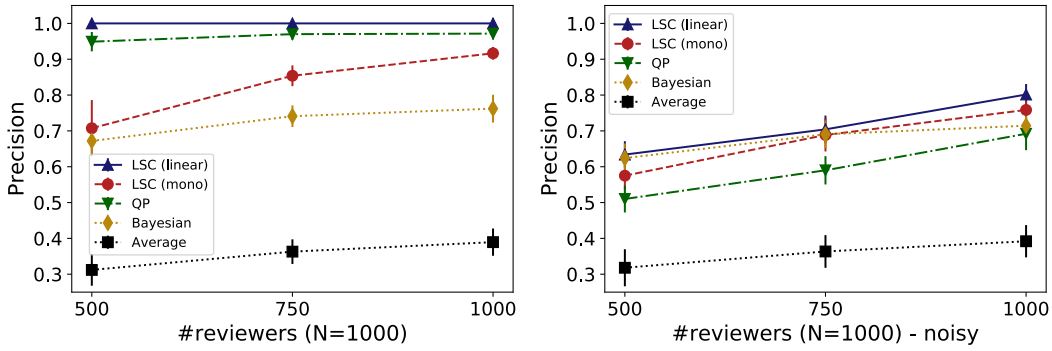

Figure 5: The experiments evaluate the performance of baseline models as well as the LSC under monotone or linear constraints. The first plot shows how the algorithms' performance changes according to $N : M$, **the paper to reviewer ratio** in the **noiseless** $(\sigma = 0)$ setting, the second plot shows the performance change in the **noisy** $(\sigma = 0.5)$ setting.

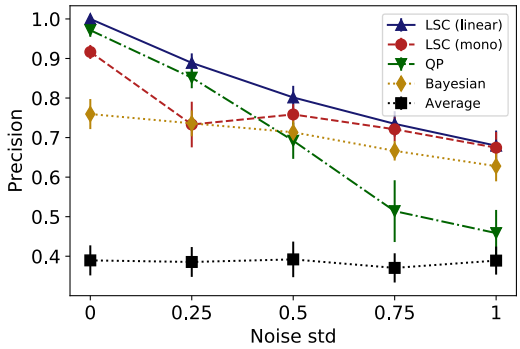

Figure 6: The experiments evaluate the performance of baseline models as well as the LSC under monotone or linear constraints. The plot shows how the algorithms' performance changes according to $sigma$, **the noise scale**.

factor for the performance. Our model LSC (linear) can always perfectly recover the true qualities of the best papers under the noiseless case even when the reviewer to paper ratio is $1 : 2$.

The noise scale σ All the previous empirical studies are to understand the impact from the structure of the *review graph*. Finally, we explore a different dimension that adds difficulty to calibration, i.e., the noise scale of perception. Results are shown in Figure 6. We can see that while the performance our proposed calibration model degrades gracefully as the noise scale increases, it steadily outperforms all the baseline methods. Among them, the Bayesian model is most robust against the noise, while the QP algorithm is very sensitive to noise. Moreover, as noise scale increases, the advantage of LSC (linear) over LSC (mono) gradually disappears. This is because large noise level makes the prior knowledge of linear scoring function less useful, as we have pointed out in Appendix B.3.

B.5 Performance Comparison in Rank-Aware Metrics

In this section, we further investigate the quality of our calibration through rank-aware metrics, that are typically used in the evaluation of information retrieval and recommendation system. Specially, we consider the following metrics that are designed to measure different aspects of ranking:

- **Average L1** $\omega(\mathbf{x}; \mathbf{x}^*) = \frac{1}{N} \sum_{i \in [N]} |\text{Rank}(i|\mathbf{x}^*) - \text{Rank}(i|\mathbf{x})|$ where $\text{Rank}(i|\mathbf{x})$ is the rank of paper i under scores $\mathbf{x}$. It measures the L1 distance of each item's ranking given respectively by the true qualities and the recovered qualities. It is used to quantify how close the recovered ranking

is to the true ranking. The smaller this distance is, the closer the recovered ranking is to the ranking of true qualities.

- **Average Precision (AP)** $\phi(S,T) = \frac{1}{p(|S|)} \sum_{i \in |S|} p(i) \cdot \mathbf{1}[i \in T]$ where $p(i)$ is the number of papers that are in T and ranked at least as high as i in S. It measures if the items in T are indeed recovered with relatively high quality by the model as the accepted papers. The larger this metric is, the better.
- **Normalized Discounted Cumulative Gain (NDCG)** $\psi(S,T) = \frac{\sum_{i \in |S|} \log^{-1}(\text{Rank}(i|\mathbf{x})+1) \cdot \mathbf{1}[i \in T]}{\sum_{i \in |S|} \log^{-1}(\text{Rank}(i|\mathbf{x})+1)}$. Similar to AP, it measures the ranking quality of the items that are both in S and T. In particular, the smooth logarithmic discounting factor has a good theoretical basis [28]. The larger this metric is, the more top papers in T are recovered with high qualities in S.

Note that the average L1 metric does not apply to Bayesian model, because it is designed to estimate the likelihood of each paper getting accepted, and thus cannot provide reasonable quality estimation of the unaccepted papers. In Table 4, we can see that our proposed LSC model with linear constraint still have the best performance in the three different ranking metrics. Meanwhile, the performance of QP in ranking metrics under the noiseless setting is close to perfect, which suggests that this baseline only accepts very few papers that should be rejected and ranked them relatively low among the accepted papers. The Bayesian model however is more robust, as its performance drops the least from noiseless to noisy setting.

Model \ Metric	Avg. L1	AP (%)	NDCG (%)	Avg. L1	AP (%)	NDCG (%)
Average	209.7 ± 6.2	60.8 ± 7.9	45.6 ± 4.4	206.6 ± 6.6	61.1 ± 6.6	45.8 ± 4.4
QP	5.5 ± 2.2	99.8 ± 0.4	97.9 ± 1.2	79.7 ± 14.2	74.2 ± 9.0	68.9 ± 6.9
Bayesian	N/A	87.1 ± 5.7	75.9 ± 4.3	N/A	83.2 ± 5.6	71.4 ± 4.0
LSC (mono)	34.4 ± 3.0	99.4 ± 0.3	93.9 ± 1.4	69.0 ± 2.8	89.1 ± 3.7	79.2 ± 2.4
LSC (linear)	$\mathbf{0 \pm 0}$	$\mathbf{100 \pm 0}$	$\mathbf{100 \pm 0}$	$\mathbf{54.5 \pm 2.0}$	$\mathbf{95.4 \pm 1.6}$	$\mathbf{84.7 \pm 2.3}$

Table 4: Experimental results for **linear scoring functions** setting on the Average, QP [21], Bayesian [10], and LSC with monotone, and linear constraints. Each entry contains the mean and standard deviation over 20 trials. The table on the left and right side respectively shows the results in the **noiseless** ($\sigma = 0$) and **noisy** ($\sigma = 0.5$) setting.

B.6 More results on Peer-Grading Dataset

In this section, we include more results on the Peer-Grading Dataset. Each homework has about 250 submissions and 200 student reviewers; each submission have at least 6 reviews. However, due to the different settings and difficulty levels of each homework, the calibration results are slightly different. We here include all of these results in Figure 7.

We can see that the performance of average baseline is generally good in the Precision metric, but bad in the ranking metric. This suggests it is important to dig into the ranking metric on how the recovered quality can match the ground truth order. Each baseline seems to have certain scenario that they can top the remaining models. And this potentially means our models designed for peer reviews may not be sufficient to model all the factors in the peer grading tasks. However, to our best effort, this is the only real-world dataset to test our model performance beyond synthesized dataset. And it is fair to conclude that LSC (linear) does show the overall best and most stable calibration performance in different ranking metrics for different n.

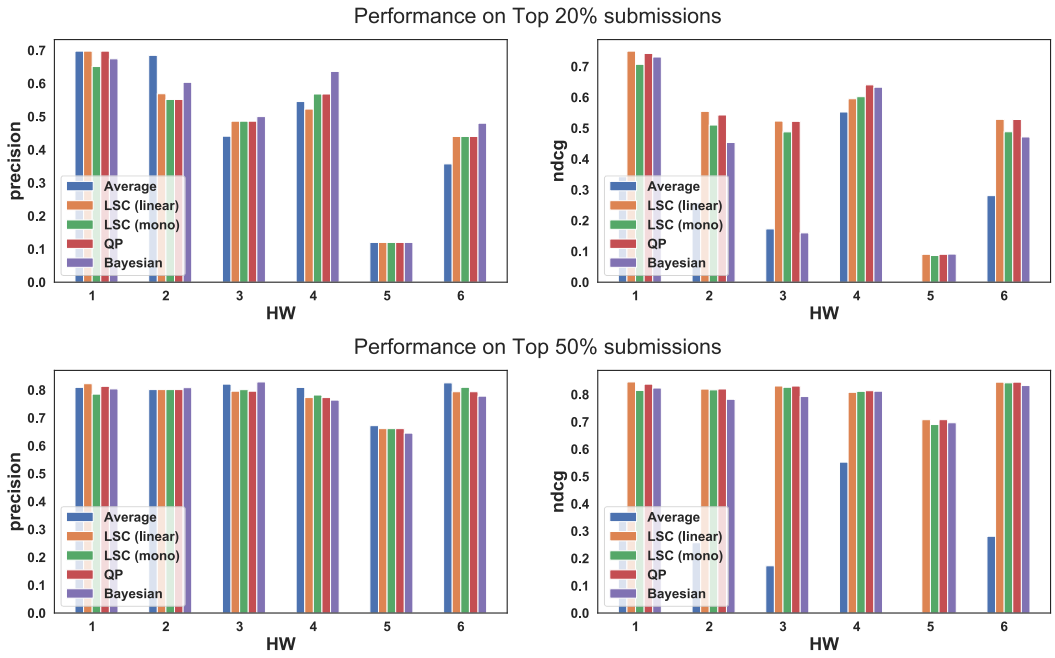

Figure 7: Experimental results on Peer-Grading dataset. In each plot, the performance of each model is compared in six different homework. n is set as the 20%, 50% of total papers to be selected in the plots of each row from top to bottom. We use the metric, Precision, NDCG in the plots of each column from left to right.