Counterfactual Evaluation of Peer-Review Assignment Policies
Supplemental Material

Martin Saveski, Steven Jecmen, Nihar B. Shah, Johan Ugander

A Linear Programs for Peer-Review Assignment

Deterministic Assignment. Let $Z \in \{0,1\}^{|\mathcal{R}|\times|\mathcal{P}|}$ be an assignment matrix where $Z_{r,p}$ denotes whether reviewer $r \in \mathcal{R}$ is assigned to paper $p \in \mathcal{P}$. Given a matrix of similarity scores $S \in \mathbb{R}_{\geq 0}^{|\mathcal{R}|\times|\mathcal{P}|}$, a standard objective is to find an assignment of papers to reviewers that maximizes the sum of similarities of the assigned pairs, subject to constraints that each paper is assigned to an appropriate number of reviewers ℓ, each reviewer is assigned no more than a maximum number of papers k, and conflicts of interest are respected [5, 27, 30–34]. Denoting the set of conflict-of-interest pairs by $\mathcal{C} \subset \mathcal{R} \times \mathcal{P}$, this optimization problem can be formulated as the following linear program:

$$\begin{aligned}
\max_{Z_{r,p}: r\in\mathcal{R}, p\in\mathcal{P}} \quad & \sum_{r\in\mathcal{R}, p\in\mathcal{P}} Z_{r,p} S_{r,p} & \\
\text{s.t.} \quad & \sum_{r\in\mathcal{R}} Z_{r,p} = \ell & \forall p \in \mathcal{P} \\
& \sum_{p\in\mathcal{P}} Z_{r,p} \leq k & \forall r \in \mathcal{R} \\
& Z_{r,p} = 0 & \forall (r,p) \in \mathcal{C} \\
& 0 \leq Z_{r,p} \leq 1 & \forall r \in \mathcal{R}, p \in \mathcal{P}.
\end{aligned}$$

By total unimodularity conditions, this problem has an optimal solution where $Z_{r,p} \in \{0,1\}, \forall r \in \mathcal{R}, p \in \mathcal{P}$.

Although the above strategy is the primary method used for paper assignments in large-scale peer review, other variants of this method have been proposed and used in the literature. These algorithms consider various properties in addition to the total similarity, such as fairness [35, 36], strategyproofness [37, 51], envy-freeness [47] and diversity [52]. We focus on the sum-of-similarities objective here, but our off-policy evaluation framework is agnostic to the specific objective function.

Randomized Assignment. As one approach to strategyproofness, Jecmen et al. [16] introduce the idea of using randomization to prevent colluding reviewers and authors from being able to guarantee their assignments. Specifically, the algorithm computes a randomized paper assignment, where the marginal probability $P(Z_{r,p})$ of assigning any reviewer r to any paper p is at most a parameter $q \in [0,1]$, chosen a priori by the program chairs. These marginal probabilities are determined by the following linear program, which maximizes the expected similarity of the assignment:

$$\begin{aligned}
\max_{P(Z_{r,p}): r\in\mathcal{R}, p\in\mathcal{P}} \quad & \sum_{r\in\mathcal{R}, p\in\mathcal{P}} P(Z_{r,p}) S_{r,p} & \\
\text{s.t.} \quad & \sum_{r\in\mathcal{R}} P(Z_{r,p}) = \ell & \forall p \in \mathcal{P} \\
& \sum_{p\in\mathcal{P}} P(Z_{r,p}) \leq k & \forall r \in \mathcal{R} \\
& P(Z_{r,p}) = 0 & \forall (r,p) \in \mathcal{C} \\
& 0 \leq P(Z_{r,p}) \leq q & \forall r \in \mathcal{R}, p \in \mathcal{P}.
\end{aligned} \tag{3}$$

A reviewer-paper assignment is then sampled using a randomized procedure that iteratively redistributes the probability mass placed on each reviewer-paper pair until all probabilities are either zero or one. This procedure ensures only that the desired marginal assignment probabilities are satisfied, providing no guarantees on the joint distributions of assigned pairs.

B "No Interference" Assumption

Our estimators assume that there is no interference between the units, i.e., that the treatment of one unit does not affect the outcomes for the other units. In the causal inference literature, this assumption

is referred to as the Stable Unit Treatment Value Assumption (SUTVA) [53]. In the context of peer review, SUTVA implies that: (i) The quality $Y_{r,p}$ of the review by reviewer r reviewing paper p does not depend on what other reviewers are assigned to paper p; and (ii) the quality also does not depend on the other papers that reviewer r was assigned to review. The first assumption is quite realistic as in most peer review systems the reviewers cannot see other reviews until they submit their own. The second assumption is important to understand, as there could be "batch effects": a reviewer may feel more or less confident about their assessment (if measuring quality by confidence) depending on what other papers they were assigned to review. We do not test for batch effects or other violations of SUTVA in this work, which typically require either strong modeling assumptions or complex experimental designs [54–56] specifically tailored for testing SUTVA, but consider it important future work.

C Variance Computation and Confidence Interval Construction

In this section, we discuss how we construct confidence intervals around μ_B under Manski, monotonicity, and Lipschitz assumptions. To construct confidence intervals, we estimate the variance of $\widehat{\mu}_B(Y^{\textit{Impute}})$ as follows:

$$\widehat{\text{Var}}[\widehat{\mu}_B(Y^{\textit{Impute}})] = \frac{1}{N^2} \sum_{(i,j)\in(\mathcal{R}\times\mathcal{P})^2} \text{Cov}[Z_i, Z_j] Z_i^A Z_j^A W_i W_j Y_i' Y_j',$$

$$\text{where } Y_i' = \begin{cases} Y_i & \text{if } i \in \mathcal{I}^+ \\ Y_i^{\textit{Impute}} & \text{if } i \in \mathcal{I}^{Att} \cup \mathcal{I}^- \\ \overline{Y} & \text{if } i \in \mathcal{I}^{Abs}. \end{cases}$$

The covariance terms (taken over $Z \sim P_A$) are not known exactly, owing to the fact that the procedure by Jecmen et al. [16] only constrains the marginal probabilities of individual reviewer-paper pairs, but pairs of pairs can be non-trivially correlated. In the absence of a closed-form expression, we use Monte Carlo methods to tightly estimate these covariances. In both our analyses of the TPDP and AAAI datasets, we sampled 1 million assignments and computed the empirical covariance. We ran an additional analysis to investigate the variability of our variance estimates. We took a bootstrap sample of 100,000 assignments (from the set of all 1 million assignments we sampled) and computed the variance based only on the (smaller) bootstrap sample. We repeated this procedure 1,000 times and computed the variance of our variance estimates. We found that the variance of our variance estimates is very small (less than 10^{-9}) even when we use 10 times fewer sampled assignments, suggesting that we have sampled enough assignments to accurately estimate the variance.

When employing Manski bounds, we can adapt a well-established inference procedure to construct 95% confidence intervals of a partial identification region that asymptotically contain the true value of μ_B with probability at least 95%. Following Imbens and Manski [43], we construct the interval:

$$\widehat{\mu}_B^{CI} \in \left[\widehat{\mu}_B(y_{\min}) - z'_{\alpha,n}\sqrt{\widehat{\text{Var}}[\widehat{\mu}_B(y_{\min})]/N},\ \widehat{\mu}_B(y_{\max}) + z'_{\alpha,n}\sqrt{\widehat{\text{Var}}[\widehat{\mu}_B(y_{\max})]/N}\right],$$

where the z-score analog $z'_{\alpha,n}$ ($\alpha = 0.95$), is set by their procedure such that the interval asymptotically has at least 95% coverage under plausible regularity conditions; for further details, see the discussion below. When employing the monotonicity assumption, we construct the interval:

$$\widehat{\mu}_{B|M}^{CI} \in \left[\widehat{\mu}_B(\widetilde{Y}_{min}^{Mon}) - z'_{\alpha,n}\sqrt{\widehat{\text{Var}}[\widehat{\mu}_B(\widetilde{Y}_{min}^{Mon})]/N},\ \widehat{\mu}_B(\widetilde{Y}_{max}^{Mon}) + z'_{\alpha,n}\sqrt{\widehat{\text{Var}}[\widehat{\mu}_B(\widetilde{Y}_{max}^{Mon})]/N}\right],$$

and when employing the Lipschitz assumption, we construct the interval:

$$\widehat{\mu}_{B|L}^{CI} \in \left[\widehat{\mu}_B(\widetilde{Y}_{min}^{L}) - z'_{\alpha,n}\sqrt{\widehat{\text{Var}}[\widehat{\mu}_B(\widetilde{Y}_{min}^{L})]/N},\ \widehat{\mu}_B(\widetilde{Y}_{max}^{L}) + z'_{\alpha,n}\sqrt{\widehat{\text{Var}}[\widehat{\mu}_B(\widetilde{Y}_{max}^{L})]/N}\right].$$

These intervals converge uniformly to the specified α-level coverage under a set of regularity assumptions on the behavior of the estimators of the upper and lower endpoints of the interval estimate: Assumption 1 from [43], establishing the coverage result in Lemma 4 there. It is difficult to verify whether Assumption 1 is satisfied for the designs (sampling reviewer-paper matchings) and interval endpoint estimators (Manski, monotonocity, Lipschitz) in this work.

A different set of assumptions, most significantly that the fraction of missing data is known before assignment, support a different method for computing confidence intervals with the coverage result in Lemma 3 from [43], obviating the need for Assumption 1. In our setting, small amounts of attrition (relative to the number of policy-induced positivity violations) mean that the fraction of data that is missing is not exactly known before assignment, but almost. In practice, we find that the Imbens-Manski interval estimates from their Lemma 3 (assuming a known fraction of missing data) and Lemma 4 (assuming Assumption 1) are nearly identical for all three of the Manksi-, monotonicity-, and Lipschitz-based estimates, suggesting the coverage is well-behaved. A detailed theoretical analysis of whether the estimators obey the regularity conditions of Assumption 1 is beyond the scope of this work.

D Linear Programs for Partial Identification Based on Lipschitz Smoothness

In this section, we provide a more detailed description of the linear programs used to compute our partial identification estimates assuming Lipschitz smoothness on the correspondence between covariates and outcomes. Intuitively, the Lipschitz smoothness assumption captures the idea that we expect two reviewer-paper pairs that are very similar in covariate space to have similar expertise. For covariate vectors X_i and X_j, define $d(X_i, X_j)$ as some notion of distance between the covariates. Then, the Lipschitz assumption states that there exists a constant L such that $|Y_i - Y_j| \leq Ld(X_i, X_j)$ for all i and j in $\mathcal{R} \times \mathcal{P}$.

As in our formulation for partial identification based on monotonicity (Section 4, LP (2)), we introduce surrogate values $\widetilde{Y}_i$ and implement a two-level optimization problem to address Lipschitz violations within the observed outcomes, i.e., if two observed pairs are very close in covariate space but have different outcomes. We also define the universe of relevant pairs $\mathcal{U} = \mathcal{O} \cup \mathcal{I}^{Att} \cup \mathcal{I}^{Abs} \cup \mathcal{I}^{-}$ and a very large constant Ψ.

This results in the following pair of optimization problems:

$$
\begin{aligned}
(\widetilde{Y}^L_{min}, \widetilde{Y}^L_{max}) = \underset{\widetilde{Y}_i : i \in \mathcal{U}}{\operatorname{argmin}} \; & \Psi \sum_{i \in \mathcal{O}} |\widetilde{Y}_i - Y_i| \pm \left(\sum_{i \in \mathcal{I}^{Att} \cup \mathcal{I}^{Abs}} \widetilde{Y}_i W_i + \sum_{i \in \mathcal{I}^{-}} \widetilde{Y}_i P_B(Z_i) \right) \\
\text{s.t.} \quad & |\widetilde{Y}_i - \widetilde{Y}_j| \leq Ld(X_i, X_j) \quad \forall (i,j) \in \mathcal{U}, \\
& y_{min} \leq \widetilde{Y}_i \leq y_{max} \qquad\qquad \forall i \in \mathcal{U}.
\end{aligned}
$$

The sign of the second objective term depends on whether a lower (negative) or upper (positive) bound is being computed. The last set of constraints are the same constraints used to construct the Manski bounds described in Section 4, which here are combined with the Lipschitz assumption to jointly constrain the possible outcomes. In the limit, as $L \to \infty$, the Lipschitz constraints become vacuous and we recover the Manski bounds. This problem can be reformulated and solved as a linear program using standard techniques.

E Model Implementation

To impute the outcomes of the unobserved reviewer-paper pairs, we train classification, ordinal regression, and collaborative filtering models. Classification models are suitable since the reviewers select their expertise and confidence scores from a set of pre-specified choices. Ordinal regression models additionally model the fact that the scores have a natural ordering. Collaborative filtering models, in contrast to the classification and ordinal regression models, do not rely on covariates and instead model the structure of the observed entries in the reviewer-paper outcome matrix, which is akin to user-item rating matrices found in recommender systems.

In the classification and regression models, we use the covariates X_i for each reviewer-paper pair as input features. In our analysis, we consider the two/three component scores used to compute the similarities: for TPDP, $X_i = (T_i, B_i)$; for AAAI, $X_i = (T_i, K_i, B_i)$. These are the primary components used by conference organizers to compute similarities, so we expect them to be usefully correlated with match quality. Although we perform our analysis with this choice of covariates, one

Figure 2: Test performance of the imputation models described in Section 4 (Model Imputation), averaged across 10 random train/test splits of all observed reviewer-paper pairs. The error bars show 95% confidence intervals.

could also include various other features of each reviewer-paper pair, e.g., some encoding of reviewer and paper subject areas, reviewer seniority, etc.

To evaluate the performance of the models, we randomly split the observed reviewer-paper pairs into train (75%) and test (25%) sets, fit the models on the train set, and measure the mean absolute error (MAE) of the predictions on the test set. To get more robust estimates of the performance, we repeat this process 10 times. In the training phase, we use 10-fold cross-validation to tune the hyperparameters, using MAE as a selection criterion, and retrain the model on the full training set with the best hyperparameters. We also consider two preprocessing decisions: (*a*) whether to encode the bids as one-hot categorical variables or continuous variables with the values described in Section 5, and (*b*) whether to standardize the features. In both cases, we used the settings that, overall, worked best (at prediction) for each model. We tested several models from each model category. To simplify the exposition, we only report the results of the two best-performing models in each category. The code repository referenced in Section 1 contains the implementation of all models, including the sets of hyperparameters considered for each model.

Figure 2 shows the test MAE across the 10 random train/test splits (means and 95% CIs) using expertise and confidence outcomes for both TPDP and AAAI. We note that all models perform significantly better than a baseline that predicts the mean outcome in the train set. For TPDP, we find that all models perform similarly, except for *cf-svd++*, which performs slightly better than the other models, both for expertise and confidence. For AAAI, all classification and regression models perform similarly, but the collaborative filtering models perform slightly worse. This difference in performance is perhaps due to the fact that we consider a larger set of covariates for AAAI than TPDP, likely making the classification and ordinal regression models more predictive.

Finally, to estimate $\widehat{\mu}_B$, we train each model on the set of all observed reviewer-paper pairs, predict the outcomes for all unobserved pairs, and impute the predicted outcomes as described in Section 4. In the training phase, we use 10-fold cross-validation to select the hyperparameters and refit the model on the full set of observed reviewer-paper pairs.

F Details of AAAI Assignment

In Section 5, we described a simplified version of the stage of the AAAI assignment procedure that we analyze, i.e., the assignment of senior reviewers to the first round of submissions. In this section, we describe this stage of the AAAI paper assignment more precisely.

A randomized assignment was computed between 3145 senior reviewers and 8450 first-round paper submissions, independent of all other stages of the reviewer assignment. The set of senior reviewers was determined based on reviewing experience and publication record; these criteria were external to the assignment. Each paper was assigned $\ell = 1$ senior reviewer. Reviewers were assigned to at

most $k = 4$ papers, with the exception of reviewers with a "Machine Learning" primary area or in the "AI For Social Impact" track, who were assigned to at most $k = 3$ papers. The probability limit was $q = 0.52$.

The similarities were computed from text-similarity scores T_i, subject-area scores K_i, and bids B_i. Either the text-similarity scores or the area scores could be missing for a given reviewer-paper pair, due to either a reviewer failing to provide the needed information or due to other errors in the computation of the scores. The text-similarity scores T_i were created using text-based scores from two different sources: (i) the Toronto Paper Matching System (TPMS) [27], and (ii) the ACL Reviewer Matching code [28]. The text-similarity scores T_i was set equal to the TPMS score for all pairs where this score was not missing, set equal to the ACL score for all other pairs where the ACL score was not missing, and marked as missing if both scores were missing. The subject-area scores were computed from reviewer and paper subject areas using the procedure described in Appendix A of [6].

Next, base scores $S'_i = w_{\text{text}} T_i + (1 - w_{\text{text}}) K_i$ were then computed with $w_{\text{text}} = 0.75$, if both T_i and K_i were not missing. If either T_i or K_i was missing, the base score was equal to the non-missing score of the two. If both were missing, the base score was set as $S'_i = 0$. For pairs where the bid was *"willing"* or *"eager"* and $K_i = 0$, the base score was set as $S'_i = T_i$.

Next, final scores were computed as $S_i = {S'_i}^{1/B_i}$, using the bid values *"not willing"* (0.05), *"not entered"* (1), *"in a pinch"* $(1+0.5\lambda_{\text{bid}})$, *"willing"* $(1+1.5\lambda_{\text{bid}})$, *"eager"* $(1+3\lambda_{\text{bid}})$; with $\lambda_{\text{bid}} = 1$. If $S_i < 0.15$ and K_i was not missing, the final score was recomputed as $S_i = \min(K_i^{1/B_i}, 0.15)$. Finally, for reviewers who did not provide their profile for use in conflict-of-interest detection, the final score was reduced by 10%.

In all of our analyses, we follow this same procedure to determine the assignment under alternative policies (varying only the parameters w_{text}, λ_{bid}, and q).

G Details Regarding Assumption Suitability

In this section, we provide additional details on the discussion in Section 5 on the suitability of the monotonicity and Lipschitz smoothness assumptions.

Monotonicity. Monotonicity assumes that when any component of the covariates increases, the review quality should not be lower. We can test this assumption on the observed outcomes: among all pairs of reviewer-paper pairs with both outcomes observed, 65.7% (TPDP)/28.0% (AAAI) have a dominance relationship ($X_i \succ X_j$) and of those pairs, 79.8% (TPDP)/76.4% (AAAI) satisfy the monotonicity condition. The fraction of dominant pairs for TPDP is higher since we consider only two covariates.

Lipschitz Smoothness. For the Lipschitz assumption, a choice of distance in covariate space is required. We choose the ℓ_1 distance, normalized in each dimension so that all component distances lie in $[0, 1]$, and divided by the number of dimensions. For AAAI, some reviewer-paper pairs are missing a covariate; if so, we impute a distance of 1 in that component. We then choose several potential Lipschitz constants L by analyzing the reviewer-paper pairs with observed outcomes.

First, we examine the fraction of pairs of observed reviewer-paper pairs that violate the Lipschitz condition for each value of L. Figures 3 and 4 show the CCDF of L for pairs of observations (in other words, the fraction of violating observation pairs for each value of L) with respect to expertise and confidence respectively. In our experiments in Section 5, we use values of L corresponding to less than 10%, 5%, and 1% violations from these plots.

Next, we examine the distances from unobserved reviewer-paper pairs to their closest observed reviewer-paper pair. In Figure 5, we show the CCDF of these distances for unobserved reviewer-paper pairs within a set of "relevant" pairs. We define the set of "relevant" unobserved pairs to be all pairs not supported on-policy that have positive probability in at least one policy among all off-policies with varying w_{text} with $q = 1$ for TPDP, and all off-policies varying w_{text} and λ_{bid} with $q = 1$ for AAAI.

Intuitively, the Lipschitz assumptions correspond to beliefs that the outcome does not change too much as the similarity components change. As one example, for $L = 30$ on AAAI, when one

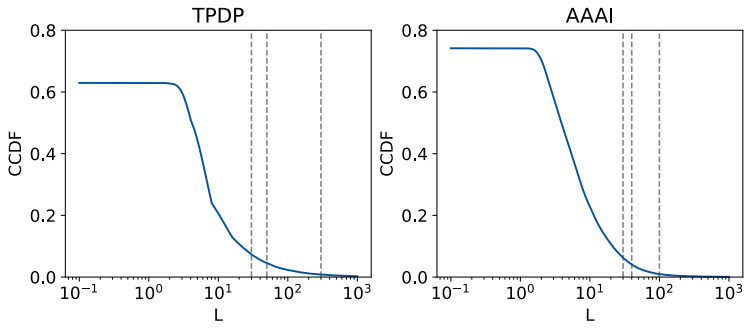

Figure 3: CCDF of the $L = |Y_i - Y_j|/d(X_i, X_j)$ values for all pairs of observed points, where Ys are *expertise* scores. The dashed lines denote the L values corresponding to less than 10%, 5%, and 1% violations. For TPDP, these values are $L = 30, 50, 300$, respectively; for AAAI, $L = 30, 40, 100$.

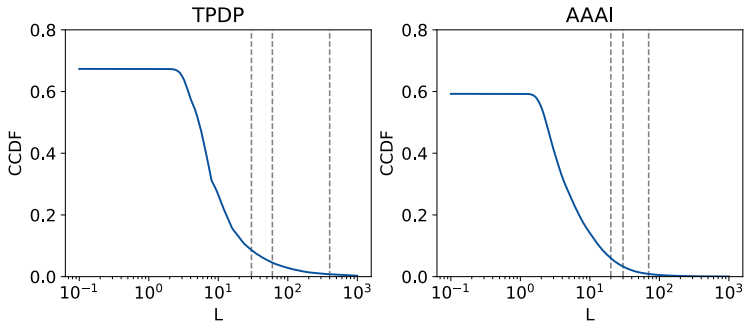

Figure 4: CCDF of the $L = |Y_i - Y_j|/d(X_i, X_j)$ values for all pairs of observed points, where Ys are *confidence* scores. The dashed lines denote the L values corresponding to less than 10%, 5%, and 1% violations. For TPDP, these values are $L = 30, 60, 400$, respectively; for AAAI, $L = 20, 30, 70$.

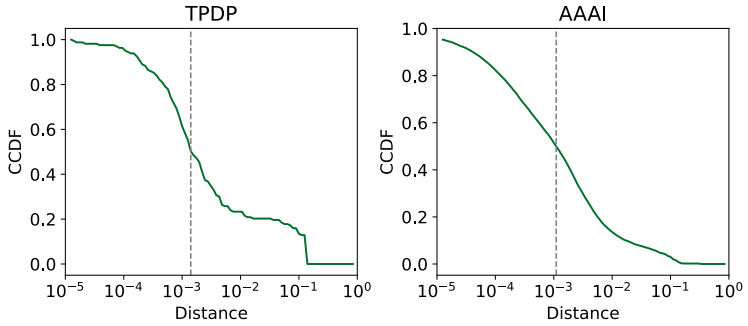

Figure 5: CCDF of the distances between each relevant unobserved reviewer-paper pair and its closest observed reviewer-paper pair. The dashed lines show the medians: 0.0014 (TPDP) and 0.0011 (AAAI).

similarity component differs by 0.1, the outcomes can differ by at most 1. Effectively, the imputed outcome of each unobserved pair is restricted to be relatively close to the outcome for the closest observed pair. From the distribution in Figure 5, we observe median distances of 0.0014 (TPDP) and 0.0011 (AAAI) across the pairs violating positivity under any of these off-policies. We conclude that most imputed pairs are very close to some observed pair, and even large values of L can significantly decrease the bound width when compared to the Manski bounds.

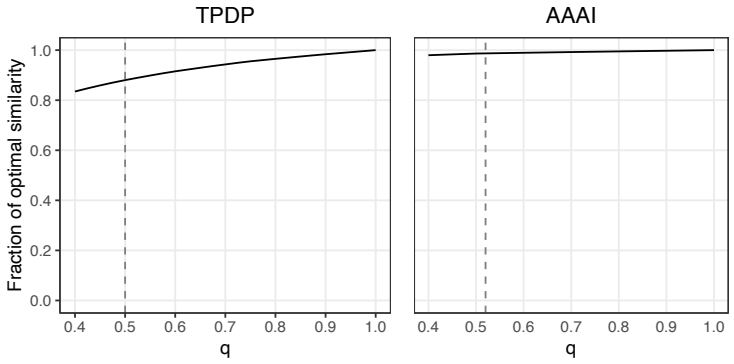

Figure 6: The "cost of randomization" as measured by the expected total assignment similarity. The plot shows the ratio between the sum of similarities under a randomized assignment (LP (3)) with ($q \leq 1$) and the sum of similarities under a deterministic assignment ($q = 1$). The dashed lines show the values of q set on-policy.

H Tie-Breaking Behavior

In Section 5 (Datasets), we specify a policy in terms of the parameters of LP (3) (specifically, by altering the parameters q, w_{text}, and λ_{bid} from the on-policy values). However, LP (3) may not have a unique solution, and thus each policy may not correspond to a unique set of assignment probabilities. Of particular concern, the on-policy specification of LP (3) does not uniquely identify the actual on-policy assignment probabilities.

Ideally, we could use the same tie-breaking methodology as was used in the on-policy to pick a solution in each off-policy to avoid introducing additional effects from variations in tie-breaking behavior. However, this behavior was not specified in the venues we analyze. To resolve this, we fixed arbitrary tie-breaking behaviors such that the on-policy solution to LP (3) matches the actual on-policy assignment probabilities; we then use these same behaviors for all off-policies.

In the TPDP analyses, we perturb all similarities by small constants such that all similarity values are unique. Specifically, we change the objective of LP (3) to $\sum_{i \in \mathcal{R} \times \mathcal{P}} P(Z_i)[(1-\lambda)S_i + \lambda \mathcal{E}_i]$, where $\lambda = 1e^{-8}$, and $\mathcal{E} \in \mathbb{R}^{|\mathcal{R}| \times |\mathcal{P}|}$ is the same for all policies. To choose $\mathcal{E}$, we sampled each entry uniformly at random from $[0, 1]$ and checked that the solution of the perturbed on-policy LP matches the on-policy assignment probabilities, resampling until it does. This value of $\mathcal{E}$ was then fixed for all policies.

In the AAAI analyses, the larger size of the similarity matrix meant that randomly choosing an $\mathcal{E}$ that recovers the on-policy solution was not feasible. Instead, we more directly choose how to perturb the similarities in order to achieve consistency with the on-policy. We change the objective of LP (3) to $\sum_{i \in \mathcal{R} \times \mathcal{P}} P(Z_i)(S_i - \epsilon \mathbb{I}[P_A(Z_i) = 0])$, where $\epsilon \in \mathbb{R}$ is chosen for each policy by the following procedure. For each policy, ϵ is chosen to be the largest value from $\{10^{-9}, 10^{-6}, 10^{-3}\}$ such that the difference in total similarity between the solution of the original and perturbed LPs is no greater than a tolerance of 10^{-5}. We confirmed that using this procedure to perturb the on-policy LP recovers the on-policy assignment probabilities, as desired.

I Similarity Cost of Randomization

In [16], Jecmen et al. empirically analyze the "cost of randomization" in terms of the expected total assignment similarity, i.e., the objective value of LP (3), as q changes. This approach is also used by conference program chairs to choose an acceptable level of q in practice. In Figure 6, we show this trade-off between q and sum-similarity (as a ratio to the optimal deterministic sum-similarity) for both TPDP and AAAI. Note that in contrast, our approach in this work is to measure assignment quality via self-reported expertise or confidence rather than by similarity. In particular, the cost of randomization for TPDP is high in terms of sum-similarity but is revealed by our analysis to be mild in terms of expertise (Section 5, Results).

Table 1: Expertise of bad policies (95% confidence intervals). $L = 50$ for TPDP and $L = 40$ for AAAI.

Policy	Manski	Monotonicity	Lipschitz
TPDP Max	[2.6115, 2.7045]	[2.6551, 2.6782]	[2.6498, 2.6744]
TPDP Min	[2.5521, 2.6126]	[2.5521, 2.5986]	[2.5521, 2.5937]
AAAI Max	[3.3919, 3.5213]	[3.4756, 3.4783]	[3.4764, 3.4809]
AAAI Min	[3.2591, 3.3846]	[3.3394, 3.3419]	[3.3396, 3.3443]

J Power Investigation: Purposefully Bad Policies

Many of the off-policy assignments we consider in Section 5 have shown to have relatively similar estimated quality. A possible explanation for this tendency is that most "reasonable" optimized policies are roughly equivalent in terms of quality, since our analyses only consider adjusting parameters of the (presumably reasonable) optimized on-policy. To investigate this possibility, we analyze a policy intentionally chosen to have poor quality.

Designing a "bad" policy that can be feasibly analyzed presents a challenge, as the on-policies are both optimized and thus rarely place probability on obviously bad reviewer-paper pairs. To work within this constraint, we look for bad policies where all reviewer pairs with zero on-policy probability are regarded as conflicts. We then contrast the deterministic ($q = 1$) policy that *maximizes* the total similarity score with the "bad" policy that *minimizes* it. Since the on-policy similarities are presumably somewhat indicative of expertise, we expect the minimization policy to be worse.

The results of this comparison are presented in Table 1. On both TPDP and AAAI, we see that our methods clearly identify the minimization policies as worse. The differences in quality between the policies becomes clearer with the addition of Lipschitz and monotonicity assumptions to address attrition. This illustrates that our methods are able to distinguish a good policy (the best of the best matches) from a clearly worse one (the worst of the best matches). Thus, it is likely that our primary analyses are simply exploring high-quality regions of the assignment-policy space, and that peer review assignment quality is often robust to the exact values of the various parameters.

K Broader Impact

We hope that our work will have a substantial positive impact on the quality of peer review processes broadly. Our methodology provides conference program chairs and other organizers with a way to evaluate the quality of alternative reviewer-paper assignments that they can easily use for post-mortem analysis. The results from such analyses can be leveraged to adjust the assignment algorithm for future iterations of the conference, ideally leading to improved review quality and improved satisfaction from both reviewers and authors.

We note two potential negative impacts that our work may have. First, our methodology requires making assumptions about the unobserved outcomes (e.g., monotonicity, Lipschitz smoothness) in order to achieve good estimates of assignment quality. If these assumptions do not hold in practice, the resulting estimates may be misleading, potentially supporting adjustments to the assignment policy that in fact decrease review quality. However, as there currently is no principled method for evaluating the quality of alternative assignment policies, the status quo is that such assignment policy choices are made based primarily on the intuition of program chairs. Our work makes the underlying assumptions explicit so that they can be directly considered.

Second, our work considers estimates of the average review quality across all assigned reviewer-paper pairs, analogous to the common sum-of-similarities objective used for review-paper assignment (see Section 2). However, past work has also considered other assignment objectives based on the fairness of the assignment [35, 36, 47, 48]. By focusing only on the average review quality, our analysis does not consider the impact on individual reviewers or papers, potentially encouraging assignment policies that produce higher overall review quality at the cost of fairness. In this work, we focused on average review quality as it is the primary metric currently used by conference organizers, but consider generalizing our approach to estimands based on other notions of fairness a very important direction for future work.

L Results for Confidence Outcomes

Figure 7 shows the results of our analyses using *confidence* as a quality measure (Y). We find that the results are substantively very similar to those reported in Section 5 using expertise.

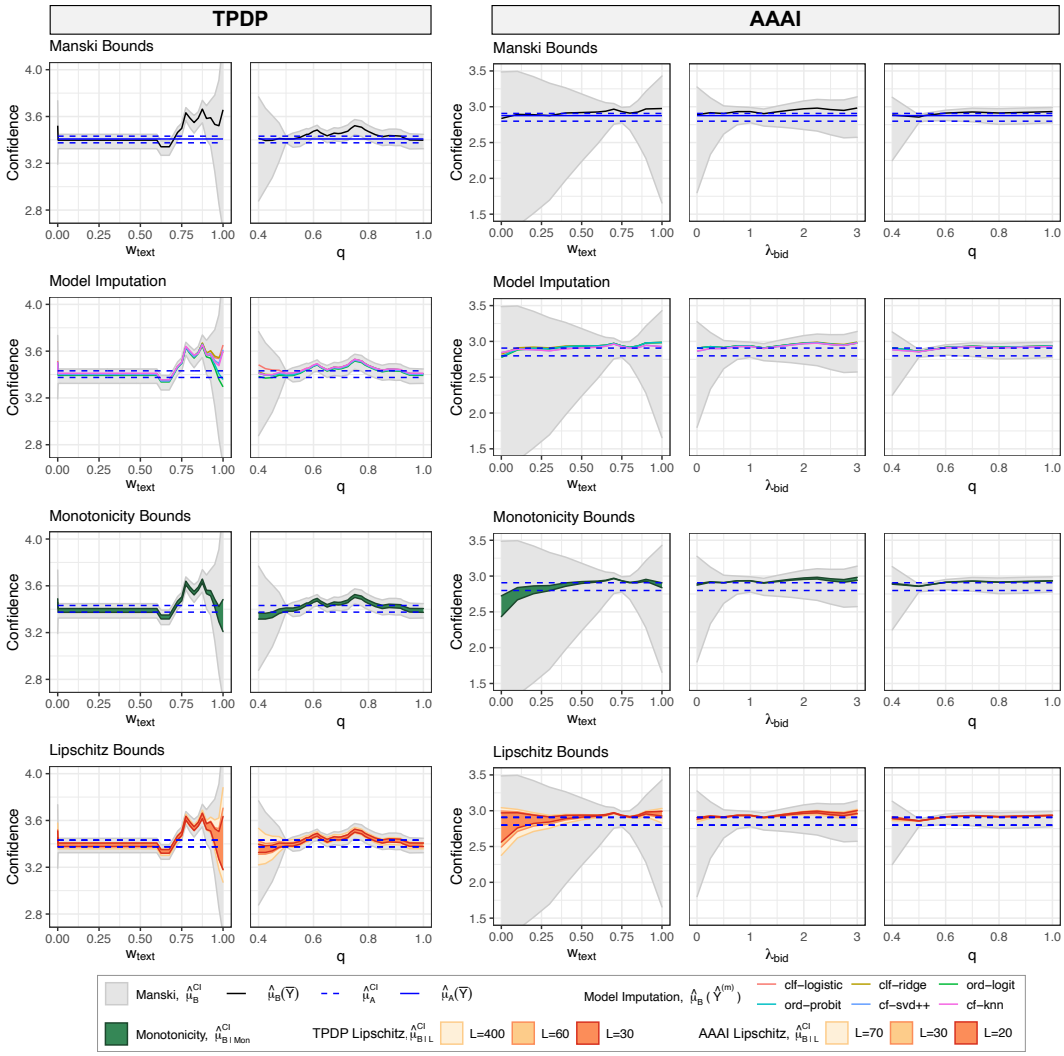

Figure 7: Confidence of off-policies varying w_{text} and q for TPDP and w_{text}, λ_{bid}, and q for AAAI, computed using the different estimation methods described in Section 4. The dashed blue lines indicate Manski bounds around the on-policy confidence, and the grey lines indicate Manski bounds around the off-policy confidence. The error bands (denoted $\widehat{\mu}_B^{CI}$ for the Manski bounds, $\widehat{\mu}_{B|M}^{CI}$ for the monotonicity bounds, and $\widehat{\mu}_{B|L}^{CI}$ for the Lipschitz bounds) represent confidence intervals that asymptotically contain the true value of μ_B with probability at least 95% as described in Appendix C. Note that to focus on the most relevant regions of the plots, the vertical axes do not start at zero.