On Testing for Biases in Peer Review

Ivan Stelmakh, Nihar Shah and Aarti Singh

Carnegie Mellon University

Tomkins, Zhang and Heavlin (2017)

Find biases in single blind setup

Their test has issues

We propose a fix

Analysis of prior work

Statistical test

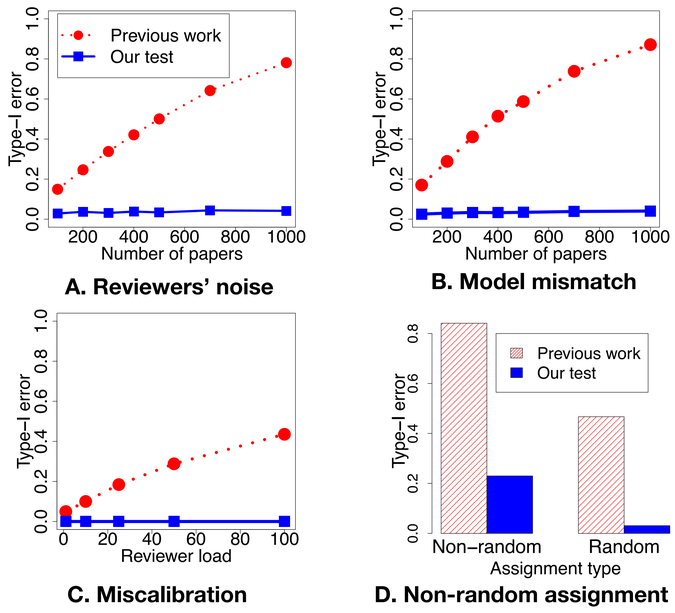
- **Objective score model.** Each paper has «true» underlying quality
- **Logistic model.** Strict parametric model of reviewers' behaviour
- **DB reviewers as estimators.** DB reviewers estimate true qualities of papers
- **Wald test.** Fit accept/reject decisions of SB reviewers into the model using DB estimates and apply standard test

Negative results. Limitations

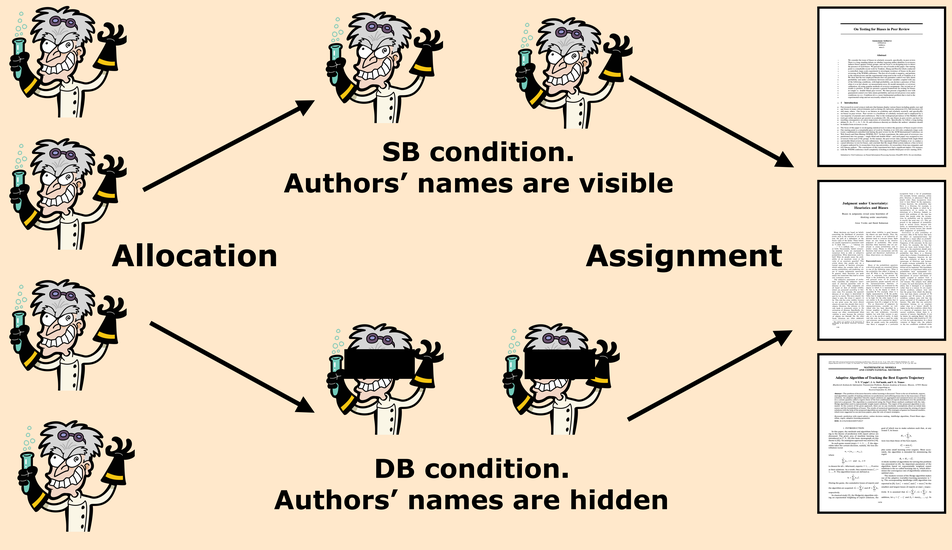
- **Humans are complex.** Parametric logistic model is unlikely to hold in practice
- **Humans are subjective.** It is known that reviewers are typically subjective
- **Humans are noisy.** DB reviewers provide noisy estimates of true scores
- **Test is specific.** Wald test relies on logistic model and may fail under small violations

Negative results. Simulations

Under reasonable conditions the test by Tomkins et al. fails to control for Type-I error rate



Setup of the experiment



Goal: test if reviewers in SB setup are biased against some categories of papers (i.e. female-authored papers)

Control over Type-I error (false positive) is of utmost importance

Disagreement test

Algorithm

- **1.** Find a set of triples (SB rev., DB rev., paper) such that each reviewer appears in at most one triple
- **2.** Condition on triples with disagreeing reviewers
- **3.** Look for trends in these triples

Our approach

Novel framework to test for biases

w_j — Protected attribute. Equals 1 iff paper's authors belong to minority category and -1 otherwise

$\pi_{ij}^{(sb)} / \pi_{ij}^{(db)}$ — Probability that reviewer i votes to accept paper j in SB/DB condition

Absence of bias. There is no difference in behaviour of SB and DB reviewers

$$H_0 : \pi_{ij}^{(sb)} = \pi_{ij}^{(db)}$$

Presence of bias. Reviewers in SB condition are more harsh (resp. lenient) to papers from minority (resp. majority) than in DB condition

$$H_1 : \begin{cases} \pi_{ij}^{(sb)} \le \pi_{ij}^{(db)} & \text{if } w_j = 1 \\ \pi_{ij}^{(sb)} \ge \pi_{ij}^{(db)} & \text{if } w_j = -1 \end{cases}$$

At least one inequality is strict

Positive result

Theorem. *The disagreement test is computationally efficient, controls for Type-I error and has non-trivial power*

Corollary. Our test is robust to issues A-C as demonstrated by simulations. Issue D is more fundamental and is tied to a setup

Impossibility result

Reviewers may behave differently in SB and DB conditions even under no bias. Can we incorporate this in the model?

Theorem. *Without assumptions on the difference in behaviour between SB and DB conditions reliable testing is impossible*

Open problems

1. Design a test and a setup s.t. setup follows standard peer review procedure and test is robust to confounding introduced by setup

2. Model the difference between SB and DB conditions and avoid impossibility result