Bayesian calibration for conference reviews

Harrison Ritz Centre for Neuroscience Studies, Queen's University

Contact: *hjr@queensu.ca* || Supported by C.V. Starr Fellowship (Princeton) and Connected Minds Fellowship (Queen's)

Abstract

Selecting among high-quality submissions is a challenge for universities, funders, and conferences.

A major difficulty in these evaluations is accounting for idiosyncrasies in reviewers' assessments.

Classical psychometric methods from Item Response Theory (IRT) can help mitigate these reviewer effects; however, these methods are rarely used in peer review.

Here, we leverage Bayesian IRT to reanalyze openly available reviews from CCN 2023. Our results suggest that IRT can support fairer assessment of conference papers and other high-stakes evaluations.

Methods

Dataset
Reviewers were asked to review a relatively large number of papers (7 or 9) on a 2D continuous scale. One axis reflected the `clarity' of the paper and the other axis reflected the `impact'. Reviewers were encouraged to use the entire scale.
[527 papers, 589 reviewers, 4808 ratings]

Model
In item-response theory (IRT), attributes are assessed using latent variable inference. These models estimate (1) the assessed value, (2) the assessor bias to give higher or lower scores("1P" models) and (3) the sensitivity of the assessor to the latent value of the assessed ("2P" models).

We fit these models using brms in R (5000 retained draws with 1000 warmup, parallelized over 16 chains). We captured clarity and impact ratings using a logit-normal model:

Each 'read' of a paper was Normally distributed with a mean ('quality') and covariance ('controversy' — disagreement between perfectly calibrated reviewers).

Reviewer generated {clarity, impact} scores by scaling the 'read' of each paper ('discrimination'), and then introducing bias ('bais') and observation noise ('noise').

'Quality' and 'controversy' were estimated for each score, and were correlated across papers (within and between scores). 'Discimination', 'bias' and 'noise' were correlated across reviewers (within and between scores).

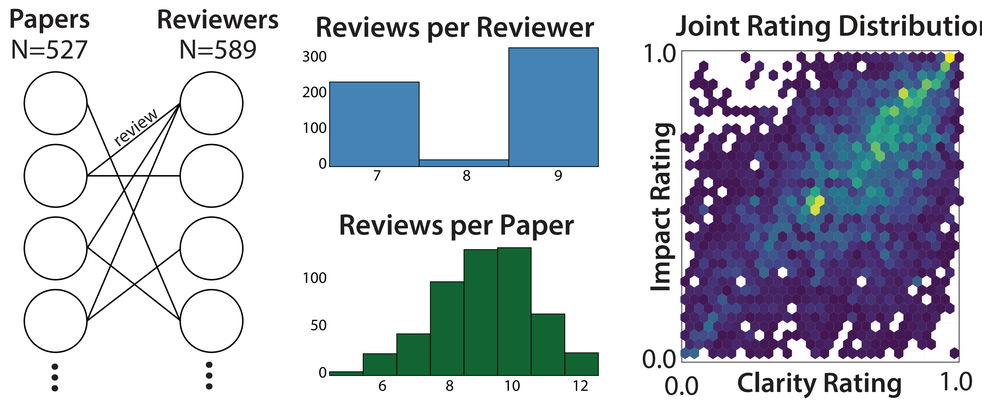
Discussion

The best-fitting model shared discrimination parameters between mean and covariance, consistent with the underlying process model.

The best-fitting model had good parameter and model recovery. We could effectively capture the distribution of scores, and distribution of emperical reviewer discrimination. This model could generalize to the new OpenReview system, though potentially at lower sensitivity due to ordinal response and lower reviewer load.

The posterior estimate from this model could help guide conference and grant decisions, such as through our proposed lottery. However, practical deployment may also want to further calibrate paper selection based on factors like topic diversity and presenter demographics.

Review Database from CCN 2023



Item Response Theory Model

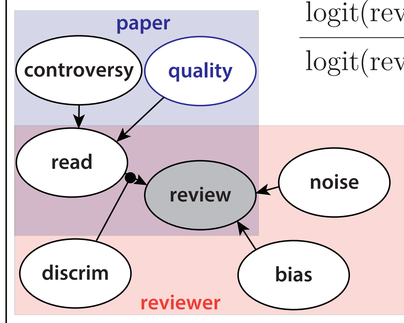
$$\text{read}_{p,r} \sim \mathcal{N}(q_p, c_p^2)$$

$$\text{logit}(\text{review}_{p,r}) \sim \mathcal{N}(d_r \text{read}_{p,r} + b_r, n_r^2)$$

$$\text{logit}(\text{review}_{p,r}) \sim \mathcal{N}(d_r q_p + b_r, d_r^2 c_p^2 + n_r^2)$$

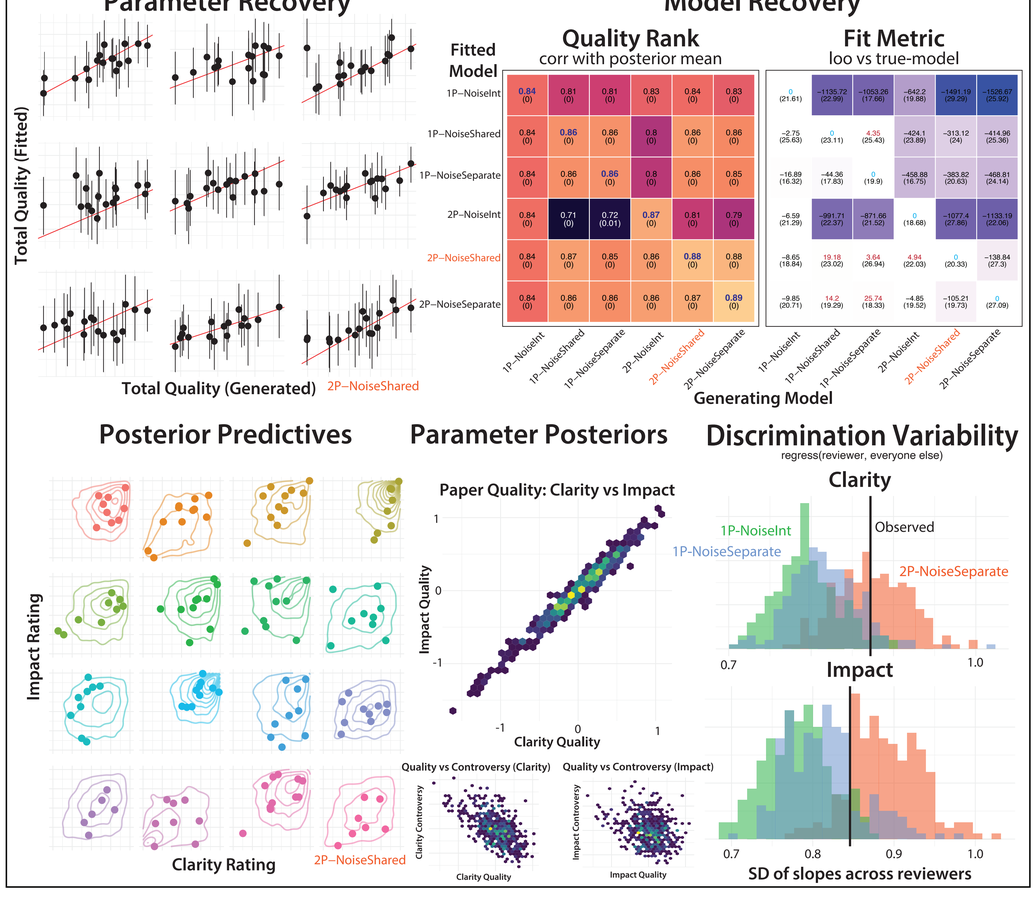
$$\begin{bmatrix} q_p \\ c_p \end{bmatrix} \sim \mathcal{N}(\mu_p, \Sigma^p)$$

$$\begin{bmatrix} d_r \\ b_r \\ n_r \end{bmatrix} \sim \mathcal{N}(\mu_r, \Sigma^r)$$

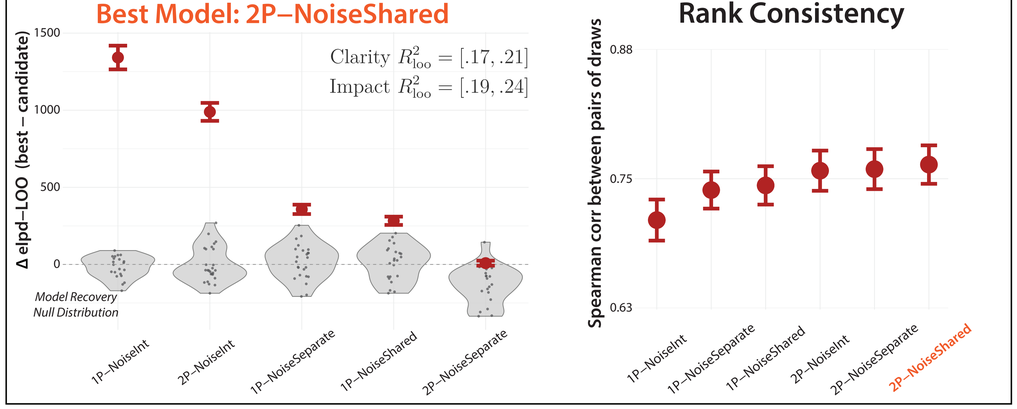


	Noise Intercept	Shared Noise Discrim	Separate Noise Discrim
Global Discrim (1P)	$\mathcal{N}(d_1 q_r + b_r, \sigma^2)$	$\mathcal{N}(d_1 q_r + b_r, d_1^2 c_p^2 + n_r^2)$	$\mathcal{N}(d_1^\mu q_r + b_r, d_1^\sigma c_p^2 + n_r^2)$
Fitted Discrim (2P)	$\mathcal{N}(d_r q_r + b_r, \sigma^2)$	$\mathcal{N}(d_r q_r + b_r, d_r^2 c_p^2 + n_r^2)$ (as above)	$\mathcal{N}(d_r^\mu q_r + b_r, d_r^\sigma c_p^2 + n_r^2)$

Model Validation



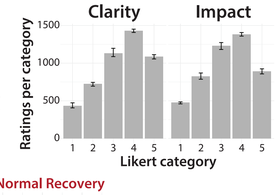
Model Comparison

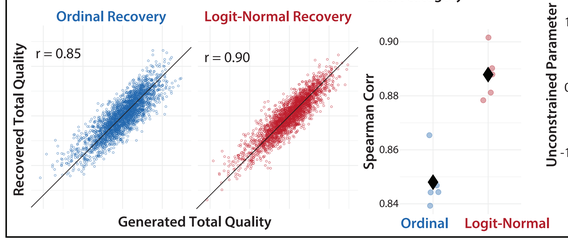


Applicability to CCN 2026

In CCN2026, reviews were submitted on the OpenReview platform, using a new ordinal rating system.

To test this approach on ordinal data, I performed parameter recovery after discretizing the posterior predictives.





CCN2026 also had a smaller reviewer load, ~2 per reviewer. In CCN2023, randomly-assigned reviewer load has a small effect on bias, and no effect on discrimination.

To simulate the effect of reviewer load, I performed parameter recovery for models that fixed reviews per paper and total number of reviews, but varied the reviews per reviewer.

Paper Selection

How can we use quality posteriors to select papers for talks?
Consider choosing 24 contributed talks, like in CCN2023.

NeurIPS approach: select papers based P(top 24).
However, there is high uncertainty, so adjacent papers are equally deserving of a talk.

Proposal: select papers with an calibrated lottery.
Randomly select papers, weighted by their exceedance probability above an self-consistent threshold (the threshold where summed exceedance equals the number of slots). Differences in P(Accept) smoothly tract difference in quality.