Title: Structured Review to Reduce Bias: Experimental Evidence on Templates, Checklists, Anonymized First-Pass Evaluation, and Gendered Outcomes in Open-Source Code Review

Keywords: code review, structured templates, checklists, anonymization, gender disparities, behavioral interventions, cognitive load, open-source software, computational social science, feminist STS, field experiment, mixed methods

Introduction

Peer code review is a cornerstone of contemporary software development and a critical locus of socio-technical gatekeeping. Yet mounting evidence suggests that ostensibly meritocratic review processes can reproduce and intensify existing inequalities, including gendered disparities in acceptance rates, comment tone, and time-to-merge. From the perspective of behavioral economics, unstructured review workflows impose cognitive demands that are unevenly absorbed by evaluators and create space for heuristic decision-making that can amplify implicit biases. From information science and feminist science and technology studies, review systems are not neutral infrastructures; their affordances construct what counts as “objective” and mediate how power circulates among maintainers and contributors. Designing structure into review processes—through templates, checklists, and procedural anonymity—offers a tractable intervention that may improve consistency and reduce bias while preserving or enhancing code quality.

This project evaluates whether structured review practices reduce gender disparities without degrading technical outcomes. We test two interventions: (1) review templates with explicit checklists embedded in the pull request and review interface; and (2) anonymized first-pass evaluation in which author identity is masked until an initial review decision or comment is submitted. We integrate a multi-site field experiment with controlled lab microtasks and qualitative interviews. The field experiment estimates causal effects on contribution outcomes; the lab component isolates cognitive mechanisms and assessor heterogeneity; interviews surface how participants interpret structure, objectivity, and power under the interventions. Together, the mixed-methods design links observed behavioral changes to underlying processes and lived experiences, informing both theory and practice.

Methods

Field experiment. We conduct a stepped-wedge, cluster-randomized trial across consenting open-source repositories hosted on major platforms. Repositories are the unit of clustering to minimize spillovers among contributors. Eligible repositories meet minimum activity thresholds (e.g., at least 15 unique contributors and 50 pull requests in the prior six months) and agree to install our open-source review bot. Prior to rollout, repositories are stratified by language ecosystem, governance model (community-led or firm-affiliated), and size, and randomized within strata to one of three sequences in a 2×2 factorial design: control, templates only, or templates plus anonymized first-pass review. The stepped-wedge schedule assigns repositories to transition from control to treatment at staggered intervals over 12 months; all clusters contribute observations under control and treatment. This design improves power and addresses time-varying confounds common in fast-moving projects.

Interventions are implemented via a bot and platform-native configuration. The template condition adds a pull request template and a review form requiring reviewers to complete a short checklist before submitting comments or decisions. Checklist items are designed with maintainers and include criteria such as test coverage, documentation updates, security implications, performance considerations, and alignment with contribution guidelines. Items are phrased behaviorally (“I verified that new code is covered by tests”) and allow structured notes. The anonymization condition masks contributor identifiers (username, avatar, affiliation badges, and contribution history) during the first-pass review. Reviewers see the diff, tests, and metadata relevant to the code but not the author. Upon submitting an initial rating or set of comments, identities are revealed for follow-up discussion and final merge decisions. Compliance is monitored via logs of checklist completion and whether initial actions occurred under masking. Repositories in control receive analytics dashboards but no new workflow components.

Primary outcomes are (i) acceptance (merge) probability; (ii) time-to-first-review and time-to-merge; and (iii) review tone. Tone is operationalized using validated text classifiers for toxicity, politeness, hedging, and constructive specificity, supplemented by manual coding on a stratified sample to calibrate thresholds and assess model error. Secondary outcomes include the number and specificity of review comments, reviewer dispersion in numeric ratings (where used), and code quality proxies: post-merge reverts, bug labels within 60 days, static analysis warnings, and CI failures at merge time. Contributor gender is measured through an opt-in, one-time survey triggered by the bot when a contributor opens a pull request, with inclusive self-identification options and the ability to opt out. Responses are linked to contributions via an encrypted token; identities are not disclosed to maintainers. Analyses that rely on gender use only records from consenting participants; sensitivity analyses test robustness to missingness patterns.

We pre-register hypotheses that structured templates and anonymized first-pass review will reduce gender gaps in acceptance, accelerate time-to-merge for underrepresented contributors, and improve tone without degrading quality. The modeling strategy uses difference-in-differences with repository and time fixed effects, interacting treatment indicators with contributor gender. Acceptance is modeled with hierarchical logistic regression; time-to-events with Cox models; tone with linear mixed effects, all with cluster-robust standard errors and random intercepts for repository and reviewer. We estimate heterogeneity by reviewer seniority, repository governance, and language ecosystem, and test for mediation by tone using causal mediation analysis. Multiple-testing adjustments control the false discovery rate. Power calculations based on historical baselines target approximately 200 repositories and 20,000 pull requests to detect a 3–5 percentage point reduction in gender gaps with 80% power.

Lab microtasks. To probe mechanisms, we run controlled review tasks with experienced developers recruited from open-source communities and professional networks. Participants are stratified by self-identified gender and seniority (years of review experience; maintainer status). In a within-subject, counterbalanced design, each participant reviews a set of synthetic pull requests with seeded defects and design trade-offs under four conditions: no template/no anonymity, template only, anonymity only, and both. Stimuli are drawn from real-world codebases and normalized for length and complexity. Outcomes include defect detection accuracy, false-positive rate, time-on-task, and inter-rater reliability. Cognitive load is measured via an adapted NASA-TLX instrument and secondary-task response latency; clickstream data capture navigation patterns. Mixed-effects models assess the impact of structure on accuracy, variance reduction, and load, and whether benefits differ by participant gender and seniority.

Interviews. Semi-structured interviews with contributors and maintainers from participating repositories (n ≈ 60–80) explore perceptions of fairness, objectivity, accountability, and the experience of masked and templated reviews. Sampling ensures representation across intervention arms, roles, and genders. Interviews are coded using an inductive–deductive approach to identify mechanisms such as perceived legitimacy of criteria, the social negotiation of checklists, and how anonymity intersects with trust and mentorship. Reflexive memos document researcher positionality; member checks validate interpretations.

Ethics and data governance. The study undergoes ethics review and uses informed consent at both repository and individual levels. Data collection minimizes personal information; gender data are optional and stored separately with strong encryption. Public release of de-identified aggregates follows privacy-preserving guidelines. Co-design workshops with maintainers shape checklist content and rollout to respect community norms.

Potential Impact

This project delivers causal evidence on whether and how structured review practices can reduce gender disparities in software development without compromising code quality. For practitioners, it provides actionable, open-source tools—templates, checklists, and an anonymized first-pass bot—along with deployment guidance tailored to different governance models. For behavioral science, it quantifies how choice architecture and reduced cognitive load affect fairness, consistency, and accuracy in real-world, high-stakes evaluation. For feminist STS and information science, it interrogates the politics of "objectivity," showing when structure reconfigures, rather than merely conceals, power in collaborative technical work.

Beyond code review, the design generalizes to other peer-evaluation settings, including academic peer review, hiring screens, and content moderation. By integrating field experimentation, controlled mechanism tests, and qualitative inquiry, the project advances methodological best practices for studying bias interventions in socio-technical systems. The resulting evidence base can inform platform policy, organizational standards, and community governance to foster more inclusive, reliable, and accountable review processes.