Proceedings of the 90th Annual International Meeting of the Psychometric Society, Minneapolis, United States (2025).

Bayesian Augmentation for Real-Time Fairness Monitoring in Assessments

Shea Valentine, *† Chris Allred, † and Amy Li †
† Harver, Dallas, TX, United States
* Corresponding author. Email: shea.valentine@libertaitech.com

Abstract
Impact ratios are an important benchmark in fairness evaluation, but they are often poorly suited to direct use as live monitoring statistics in operational settings. Repeated threshold-based significance testing is likewise poorly suited to continuous monitoring because it can behave unstably under ongoing re-evaluation. The present framework addresses this by estimating posterior subgroup pass rates, deriving divergence-based summaries, and propagating their uncertainty through an offline mapping calibrated to historical impact-ratio patterns. This yields a modular, proxy-based Bayesian framework for real-time fairness monitoring in settings where the target quantity is important but undesirable to use directly as a live deployment statistic. In sequential simulations, the proxy-based monitor closely matched the repeated-look behavior of direct posterior impact-ratio monitoring while exhibiting substantially greater stability than repeated p-valued testing.

Keywords: Bayesian, fairness, employee selection

1. Introduction

Selecting for high-performing employees is a central objective of personnel selection. Employers commonly rely on pre-hire assessments to evaluate applicant characteristics relevant to job performance and to support hiring decisions. Alongside predictive validity, selection systems must also satisfy legal and ethical requirements related to demographic fairness. Under Title VII of the Civil Rights Act of 1964, U.S. employers are prohibited from discrimination in hiring and employment decisions.

A key concern in this context is adverse impact, which occurs when a selection procedure yields meaningfully different selection rates across demographic groups. Adverse impact was articulated in Griggs v. Duke Power (Griggs v. Duke Power Co., 1971) and operationalized in the Uniform Guidelines (Equal Employment Opportunity Commission, 1978), motivating routine evaluation of subgroup selection rates. A common summary statistic used in these evaluations is the impact ratio (IR), defined as the ratio of subgroup selection rates.

Detecting adverse impact during active hiring presents substantial challenges. Once a selection procedure is deployed, changing decision rules mid-cycle undermines procedural consistency by subjecting applicants to different standards. Organizations therefore require methods that surface fairness risk without forcing immediate determinations about the acceptability of a selection system. These constraints motivate ongoing, uncertainty-aware fairness monitoring rather than discrete post hoc evaluation, especially in settings where IR remains the benchmark quantity of interest but is not used directly as the live monitoring statistic.

1. 1 Current Approaches to Evaluating Adverse Impact

Several methods are commonly used to assess adverse impact, but each has limitations for live monitoring. The four-fifths rule is a heuristic threshold that ignores sampling variability and provides no measure of uncertainty (Equal Employment Opportunity Commission, 1978). Statistical tests such as the pooled two-

sample z-test and Fisher's exact test offer formal inferential procedures, but they are designed for discrete hypothesis testing rather than repeated monitoring as evidence accrues. Effect-size measures such as Cohen's d avoid strict thresholding but remain point estimates unless supplemented with interval estimates. Taken together, these approaches emphasize discrete testing and point estimation, making them poorly suited to dynamic assessment contexts in which evidence accumulates over time.

1. 2 Bayesian Approaches to Fairness Monitoring

Bayesian methods offer an alternative approach to evaluating adverse impact by explicitly modeling uncertainty in group differences. Prior work has shown that Bayesian inference can be used to estimate credible ranges for adverse impact ratios, yielding more informative summaries than point estimates alone (Courey & Oswald, 2025). Rather than treating fairness metrics as fixed values, Bayesian approaches represent them as distributions, allowing direct probability statements about the magnitude of group disparities. Direct posterior inference for IR is therefore a natural Bayesian benchmark when the goal is to estimate IR itself.

This perspective is especially useful in applied settings with limited data, early-stage deployment, or substantial variation across selection contexts. As evidence accumulates, posterior estimates update directly, supporting ongoing fairness monitoring without reliance on sequential testing or fixed decision thresholds. In operational settings where direct use of impact ratios may be undesirable during deployment, this logic also motivates proxy-based Bayesian monitoring frameworks.

1.2.1 Implications for Organizational Practice

For organizations engaged in ongoing hiring, Bayesian frameworks support fairness monitoring that aligns with operational constraints by maintaining continuously updated uncertainty summaries rather than relying on fixed checkpoints. The present framework implements this through an offline-calibrated proxy: a fixed mapping from divergence-based summaries to historical impact-ratio behavior that is then used during monitoring. This modular design supports continuous, uncertainty-aware monitoring while keeping the benchmark quantity and the deployed live monitoring statistic separate.

1.2.2 The Present Study

Specifically, we (a) estimate posterior pass rates for demographic subgroups, (b) derive divergence-based fairness summaries from these posterior distributions, (c) propagate their uncertainty through an offline-calibrated mapping to obtain probabilistic fairness estimates across assessment contexts, and (d) evaluate the sequential monitoring behavior of this proxy-based approach relative to repeated Fisher testing and direct posterior IR monitoring under controlled simulation.

2. Methods

2. 1 Sample

The sample comprised approximately 1.5 million applicants drawn from numerous independent samples (all with n > 100) across three assessment platforms. These samples varied substantially in industry sector, geographic region, and assessment type, providing broad representation of selection contexts. For each applicant, four variables were recorded: a dichotomous pass/fail outcome, sex, age, and racioethnic group membership. Analyses of the empirical framework were conducted at the position level.

2. 2 Analytic Approach

We employed a two-stage Bayesian modeling approach in which offline calibration is separated from deployed monitoring. During calibration, observed data were used to estimate the relationship between divergence-based summaries and historical impact-ratio behavior. During monitoring, posterior uncertainty in subgroup pass rates was propagated through those summaries and the fixed calibration model to assess fairness risk over time. We do not claim that this route is more exact than direct posterior IR; rather, it defines a fixed deployable monitoring quantity for settings in which the benchmark quantity and the live monitoring statistic are intentionally kept separate.

2.2.1 Fairness Learning: Divergence-to-Fairness Mapping

For each position and demographic grouping, observed pass and fail counts were used to compute summaries of group differences, including pass-rate differences, Kullback-Leibler (KL) divergence, and Jensen-Shannon (JS) divergence (Wang et al., 2018). Using these summaries, we fit beta-regression models relating divergence patterns to historical impact-ratio behavior. Separate models were estimated for age, sex, and racioethnic groupings. These calibration models were trained offline and treated as fixed during subsequent monitoring. Historical impact ratios thus served as offline calibration targets rather than as live monitoring statistics.

2.2.2 Fairness Monitoring: Bayesian Estimation and Risk Projection

During monitoring, Bayesian models were used to estimate posterior distributions over subgroup pass rates, from which posterior distributions of divergence-based summaries were derived. Posterior samples were then propagated through the fixed offline calibration model to obtain posterior summaries of the proxy monitoring statistic, calibrated to empirical impact-ratio behavior. From this distribution, we computed posterior probabilities of fairness violations and 95% highest density intervals (HDIs) used for monitoring classification (Kruschke, 2018). Observations were classified as No Bias when the lower HDI bound exceeded .80, Definite Bias when the upper HDI bound fell below .80, and Bias Credible otherwise. This approach supports continuous, uncertainty-aware fairness monitoring without reliance on repeated threshold-based testing.

2.2.3 Implementation Details

Implementation details. Offline calibration models were fit using beta regression in Python (statsmodels), with KL divergence, JS divergence, and mean pass-rate difference as predictors. During monitoring, subgroup pass rates were modeled in PyMC with Beta priors and binomial likelihoods; posterior samples were transformed into divergence metrics and propagated through the fixed offline calibration model to obtain posterior summaries of the proxy monitoring statistic. Posterior inference was implemented in PyMC using the default NUTS sampler with 4 chains, 2,000 tuning iterations, and 5,000 retained posterior draws in total; weakly informative Beta priors were placed on subgroup pass rates.

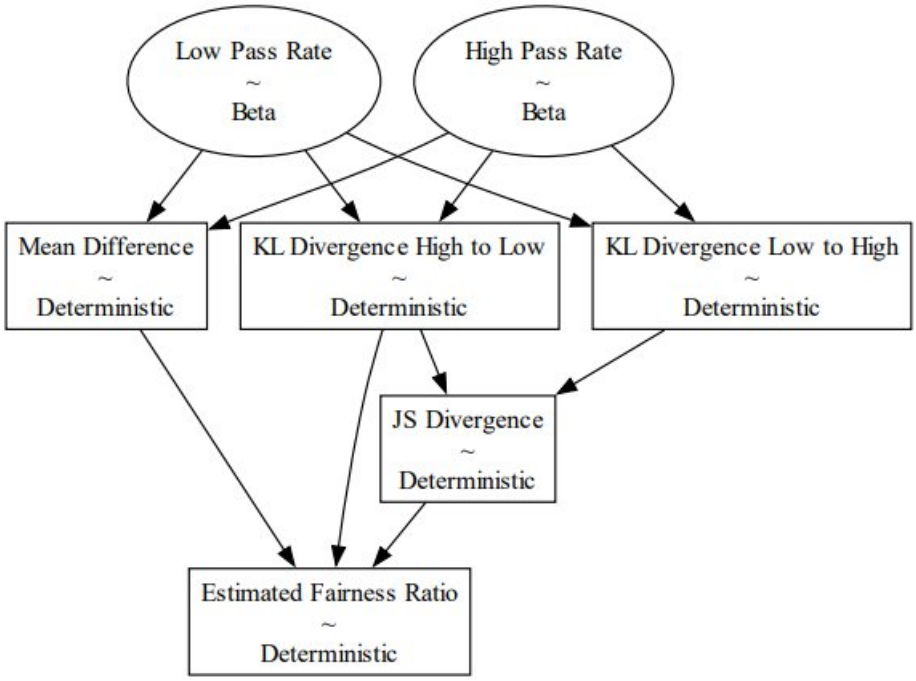

Figure 1. Graphical representation of the Bayesian model. Subgroup pass rates are modeled with Beta priors, and fairness metrics (mean difference, KL divergence, JS divergence, and the resulting fairness ratio) are deterministic functions of those parameters.

2.2.4 Sequential Monitoring Simulation Study

To evaluate the framework as a live monitoring procedure, we conducted a sequential simulation comparing repeated Fisher testing, direct posterior IR monitoring, and the proxy-based Bayesian monitor. Scenarios were defined by subgroup pass probabilities implying clearly fair, borderline, and adverse-impact conditions under balanced and imbalanced group-size settings. Monitoring began from a small initial sample and was updated one observation at a time. At each interim look, all three methods were recomputed and summarized using false-alarm rates, missed-detection rates, reversal counts, and first-alert sample size.

3. Results

3. 1 Descriptive Overview

Analyses were conducted on 255 position-level observations. The posterior mean of the proxy monitoring statistic, derived from the learned divergence-to-fairness mapping, averaged 0.80 (SD = 0.21), while the empirical impact ratio averaged 0.90 (SD = 0.10; range = 0.38-1.00).

Divergence metrics were generally small but variable. Jensen-Shannon divergence had a mean of 0.009 (SD = 0.022), and Kullback-Leibler divergence (high-to-low) averaged 0.010 (SD = 0.025). Mean pass-rate differences averaged 0.078 (SD = 0.081). Together, these distributions indicate that most assessment contexts exhibited modest group differences, with a smaller subset showing pronounced separation.

3. 2 Divergence to Fairness Mapping

Beta-regression models relating divergence metrics to empirical impact ratios demonstrated strong overall fit across demographic groupings. Regression r^2 values are reported here as descriptive checks on calibration adequacy rather than as optimization targets or as the primary contribution of the method. Posterior mean estimates of the fairness ratio showed strong correspondence with empirical impact ratios (cross-validated r^2 = .93). These calibration results are reported only to verify that the fixed proxy remained aligned with historical IR-relevant behavior; they are not offered as evidence that the proxy improves on direct posterior IR as an estimator of IR.

3. 3 Bias Classification Behavior

Posterior credible intervals were used to classify each observation into one of three monitoring categories: Definite Bias, Bias Credible, or No Bias. Table 1 presents the relationship between empirical adverse impact and estimated monitoring category. No cases with empirical adverse impact were classified as No Bias. A Fisher-Freeman-Halton exact test indicated a strong association between empirical adverse-impact status and monitoring classification (Monte Carlo $p < .0001$; $X^2(2)$ = 96.80). A follow-up Fisher's exact test on No-Bias classifications likewise supported the conclusion that the framework avoided premature declarations of fairness (two-sided p = .009).

Table 1. Empirical Adverse Impact (AI) and Estimated Bias Category

	No Bias[a]	Bias Credible	Definite Bias[b]
AI = No	114	130	3
AI = Yes	0	3	5

Table note
a No Bias indicates the lower bound of 95% HDI > 0.80.
b Definite Bias indicates the upper bound of 95% HDI < 0.80.

3. 4 Sequential Monitoring Simulation

The simulation revealed substantial differences in repeated-look behavior across methods. Fisher's exact test exhibited marked instability under continual re-evaluation, producing frequent ever-false alarms in fair and borderline settings and substantial final missed-detection rates in adverse-impact settings. For example, in the borderline balanced condition, the ever-false-alarm rate was .41 for Fisher, compared with .16 for the proxy monitor. In the moderate-adverse balanced condition, Fisher's final missed-detection rate was .60, compared with .00 for Bayesian methods. Fisher also showed higher reversal counts across repeated looks.

By contrast, the proxy-based monitor closely tracked the sequential operating behavior of direct posterior IR. Across balanced and imbalanced conditions, the two Bayesian approaches showed similar false-alarm rates, missed-detection rates, reversal counts, and first-alert sample sizes, indicating that the proxy preserved the operational behavior of posterior IR much more closely than repeated significance testing did.

3. 5. Summary

Taken together, the results support the proposed framework at two levels. First, the offline proxy showed strong calibration to empirical impact-ratio behavior and conservative empirical classification performance, including no No-Bias classifications in cases with empirical adverse impact. Second, in sequential simulation, the proxy-based monitor closely matched the operating behavior of direct posterior IR while both Bayesian monitoring approaches were substantially more stable than repeated Fisher testing. These findings support the proxy as a viable live monitoring statistic calibrated to IR-relevant behavior.

4. Discussion

This study introduces a Bayesian augmentation framework for real-time fairness monitoring in assessment-based selection systems. Rather than repeatedly testing impact-ratio benchmarks against fixed thresholds, the framework models uncertainty in subgroup pass rates and propagates that uncertainty through an offline-calibrated proxy mapping. The contribution of this approach is not that the learned mapping is more exact than direct posterior computation of the impact ratio, but that it provides a modular monitoring architecture for settings in which the target quantity is important yet undesirable to use directly during deployment. The sequential simulation results further showed that the proxy-based monitor closely tracked the repeated-look behavior of direct posterior IR while both Bayesian monitoring approaches were substantially more stable than repeated Fisher testing. In this way, the framework addresses two persistent challenges in applied fairness evaluation: instability under repeated threshold-based monitoring and the operational constraints associated with modifying selection procedures during active hiring cycles (Rouder, 2014; Gelman & Loken, 2013).

4. 1 Monitoring Rather Than Measurement

Conventional approaches to adverse-impact evaluation rely on point estimates and binary decision rules that provide limited guidance as evidence accumulates. Bayesian approaches instead represent fairness-related quantities as uncertain and update them as new data become available (Ji et al., 2020). In this paper, direct posterior IR remains the exact Bayesian benchmark for IR itself, whereas the present framework uses an offline-calibrated proxy to define a separate live monitoring statistic whose interpretation remains anchored to IR-relevant behavior.

4. 2 Conservative Behavior and Decision Safety

A central finding of this study is that the proposed monitoring framework behaves conservatively in practice. Across all position-level observations, no case was classified as no bias when empirical adverse impact was present. This directly addresses the most consequential failure mode in applied fairness monitoring: providing false reassurance in the presence of meaningful disparities.

By retaining uncertainty when evidence is equivocal and resolving toward fairness only when supported by the posterior distribution, the framework prioritizes decision safety over premature classification. The strong association between empirical adverse impact and monitoring classifications indicates that the framework remains responsive to meaningful group differences while avoiding rigid binary rules. This emphasis on conservative, uncertainty-aware assessment is consistent with recent Bayesian and probabilistic treatments of fairness metrics (Barrainkua et al., 2023).

The simulation study reinforced this conclusion in sequential settings. Relative to repeated Fisher testing, Bayesian monitoring approaches exhibited much lower repeated-look instability, including lower false-alarm rates in fair and borderline scenarios and fewer threshold reversals over time. The proxy-based monitor closely matched the operating behavior of direct posterior IR on these metrics, indicating that the proxy preserved the decision-safety profile of the Bayesian benchmark while functioning as a separate live monitoring statistic.

4. 3 Practical Implications for Fairness Monitoring

From an applied perspective, the framework offers a practical alternative to repeated impact-ratio testing during ongoing selection processes. Rather than generating abrupt decision points from repeated significance testing, it supports continuous monitoring through posterior probabilities and credibility regions. For organizations that treat IR as the legally and historically salient benchmark but prefer not to use IR itself as the deployed monitoring statistic, the present framework offers a calibrated alternative whose sequential behavior closely resembles direct posterior IR (Kruyt, 2023; Bacelar, 2021).

A natural question is why introduce a proxy when direct posterior IR is available. The reason is not inferential necessity: direct posterior IR remains the exact Bayesian benchmark for IR. Rather, the present framework addresses settings in which the benchmark quantity and the live monitoring statistic are intentionally separated, so that monitoring is conducted with a fixed offline-calibrated proxy whose interpretation remains anchored to IR-relevant behavior.

4. 4 Limitations and Future Directions

Several limitations warrant consideration. Analyses were conducted at the position level, and individual-level sample size information was not retained for all observations, limiting direct assessment of monitoring behavior as a function of raw sample size. In addition, the sequential benchmark examined only a limited set of balanced and imbalanced two-group scenarios. Future work should extend this framework to additional base-rate regimes, longer monitoring horizons, multi-group settings, hierarchical models, and alternative proxy constructions.

4. 5 Conclusion

The Bayesian framework presented here supports fairness monitoring that is uncertainty-aware, conservative, and operationally feasible. By separating offline calibration from deployed monitoring, it offers a fixed proxy monitor for settings in which the benchmark quantity and the live monitoring statistic are intentionally kept separate. In sequential simulations, the proxy closely reproduced the repeated-look behavior of direct posterior IR while both Bayesian approaches were substantially more stable than repeated Fisher testing.

Competing Interests: The framework was developed in an applied commercial context, which shaped design priorities.

References

Courey, K. A., & Oswald, F. L. (2025). Evaluating the adverse impact ratio and its associated uncertainty: A Bayesian approach. Psychology, Public Policy, and Law, 31(2), 134-146. https://doi.org/10.1037/law0000451

Equal Employment Opportunity Commission. (1978, August 25). Uniform Guidelines on Employee Selection Procedures. Federal Register, 43(166), 38290-38315 (codified at 29 C.F.R. pt. 1607). https://tile.loc.gov/storage-services/service/ll/fedreg/fr043/fr043166/fr043166.pdf

Griggs v. Duke Power Co., 401 U.S. 424 (1971). https://tile.loc.gov/storage-services/service/ll/usrep/usrep401/usrep401424/usrep401424.pdf

Ji, D., Smyth, P., & Steyvers, M. (2020). Can I trust my fairness metric? Assessing fairness with unlabeled data and Bayesian inference (arXiv:2010.09851). arXiv. https://doi.org/10.48550/arXiv.2010.09851

Barrainkua, A., Gordaliza, P., Lozano, J. A., & Quadrianto, N. (2023). Uncertainty in fairness assessment: Maintaining stable conclusions despite fluctuations (arXiv:2302.01079). arXiv. https://doi.org/10.48550/arXiv.2302.01079

Kruyt, D. M. (2023). Towards fair ML: A monitoring approach (Master's thesis, Erasmus School of Economics, Erasmus University Rotterdam). Erasmus University Thesis Repository. http://hdl.handle.net/2105/66695

Bacelar, M. O. (2021). Monitoring bias and fairness in machine learning models: A review (Version 1). ScienceOpen Preprints. https://doi.org/10.14293/S2199-1006.1.SOR-.PP59WRH.v1

Wang, H., Ustun, B., & Calmon, F. P. (2018). On the direction of discrimination: An information-theoretic analysis of disparate impact in machine learning (arXiv:1801.05398). arXiv. https://doi.org/10.48550/arXiv.1801.05398

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. Psychonomic Bulletin & Review, 21, 301-308. https://doi.org/10.3758/s13423-014-0595-4

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time (Unpublished manuscript). Columbia University. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. Advances in Methods and Practices in Psychological Science, 1(2), 270-280. https://doi.org/10.1177/2515245918771304