When Looks Make You Click "Submit": Participation-Mediated Appearance Bias in Teaching Evaluations

Anonymous Author(s)
Affiliation
Address
email

Abstract

Student teaching evaluations shape personnel decisions but are vulnerable to appearance-based biases. Using the UT Austin dataset of 463 courses taught by 94 professors, we revisit the "beauty premium" and ask a new question: *how much of the observed premium is statistically associated with student participation in the evaluation process, and when is this channel strongest?* We implement a mediation framework where perceived beauty (six independent rater scores aggregated and standardized) is correlated with evaluation participation (percent of enrolled students who submitted evaluations), which in turn is associated with mean course scores. With professor-clustered standard errors and a cluster bootstrap, we estimate that a one–standard deviation increase in beauty corresponds to a 0.099 point higher mean evaluation score on a 1–5 scale. About **17%** of this premium is statistically mediated by higher student participation (95% CI: [0.0008, 0.0354]). The direct association remains positive and significant after accounting for participation. **Surprisingly, this suggests that bias arises not only from how students rate, but also from who chooses to rate at all.** Results are descriptive (not causal), limited by potential unobserved confounding and the historical context of the dataset, but they highlight the importance of participation-driven bias when interpreting teaching evaluations.

1 Introduction

Student evaluations are widely used in hiring, promotion, and course design. A robust empirical finding is that instructors perceived as more attractive receive higher evaluations. Yet the *mechanisms* remain underexplored. One plausible and practically important channel is **participation**: attractive instructors may induce more students to fill out evaluations, shifting the composition of raters and raising the average score even without changing individual-level scoring propensities.

We study this mechanism using the well-known UT Austin dataset with six independent appearance ratings per instructor, rich instructor/course covariates, and—critically for our purpose—the *percentage of enrolled students who submitted an evaluation*. We ask: *How much of the beauty premium in mean course scores is statistically associated with evaluation participation, and when is this channel strongest?*

Contributions. (i) We quantify a participation-mediated *association* with professor-clustered inference and cluster-bootstrap CIs; (ii) we document that the mediated share is material (17%) but far from exhaustive; (iii) we show that a surprising source of bias is *who* chooses to rate, not just *how* students rate; (iv) we demonstrate heterogeneity by class size and native-English status; (v) we provide a reproducible pipeline and reporting template for evaluation policy.

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

Table 1: Beauty–Participation mediation (per 1 SD in beauty). Professor-clustered CIs.

Quantity	Estimate	95% CI
Total association with score (τ)	0.0987	[0.0076, 0.1715]
Direct association controlling for M (τ')	0.0820	[0.0052, 0.1563]
Association with participation (α_1) [pp]	2.808	[0.206, 5.355]
Association of participation with score (ϕ)	0.0060	[0.0018, 0.0095]
Product $\alpha_1\phi$ (mediated)	0.0167	[0.0008, 0.0354]
Mediated share (ratio)	0.169	[-0.010, 0.401]

2 Data and Measures

We analyze 463 course-level observations taught by 94 professors. The outcome is mean evaluation `score` (1–5). The key predictor is perceived beauty, measured by six independent rater scores; we use `bty_avg` standardized to `beauty_z`. Participation is `cls_perc_eval` (percent of enrolled students who submitted). Controls include instructor `gender`, `ethnicity`, `language` (native English), `age`, rank, course size (winsorized), and course level. Cluster identifier is `prof_id`. Cronbach's $\alpha \approx 0.90$ and a one-factor PCA confirm reliability of beauty ratings.

3 Empirical Strategy

Let X denote `beauty_z`, M the participation rate (`cls_perc_eval`), and Y the mean score. We estimate three linear models with professor-clustered standard errors:

$$\textbf{Mediator:} \quad M = \alpha_0 + \alpha_1 X + \gamma' W + u, \tag{1}$$

$$\textbf{Total association:} \quad Y = \beta_0 + \tau X + \delta' W + v, \tag{2}$$

$$\textbf{Direct association:} \quad Y = \beta_0' + \tau' X + \phi M + \delta' W + e, \tag{3}$$

where W collects controls. We interpret the product $\alpha_1\phi$ as a mediated *association*; causal interpretation requires strong assumptions (below).

3.1 Mediator Model: Fractional Logit

Because participation is a proportion, we also fit a binomial GLM with professor-clustered covariance:

$$\text{logit}(M) = \alpha_0 + \alpha_1 X + \gamma' W + u.$$

3.2 Moderated Mediation and Conditional Indirect Associations

To test heterogeneity, we allow beauty→participation and participation→score links to vary with moderators Z (class size, native-English):

$$M = \alpha_0 + \alpha_1 X + \alpha_3 XZ + \gamma' W + u, \qquad Y = \phi_0 + \phi_1 M + \phi_3 MZ + \tau' X + \delta' W + e.$$

The conditional indirect association is $(\alpha_1 + \alpha_3 Z)(\phi_1 + \phi_3 Z)$.

3.3 Interpretation and Scope

A causal mediation interpretation requires *sequential ignorability*: (i) no omitted confounders of X and M, and (ii) no omitted confounders of M and Y, conditional on W. Given potential unobservables (e.g., charisma/teaching effectiveness), we emphasize descriptive, decision-relevant associations and provide robustness checks and a sensitivity-analysis plan.

4 Results

Interpretation. Beauty is positively associated with participation (GLM $p = 0.015$), and the mediated association tends to increase with class size. For non-native instructors, estimates are imprecise with wide confidence intervals; we refrain from drawing subgroup conclusions beyond noting uncertainty.

Table 2: Robustness checks. Effects are per 1 SD increase in beauty.

Specification	Total Assoc. (τ)	Mediated (share)
Baseline (bty_avg, winsorized size)	0.0987	0.0167 (17%)
PCA latent beauty factor	0.1032	0.0159 (15%)
Trim 1% tails of class size	0.0965	0.0165 (17%)
Exclude rank/level controls	0.0948	0.0170 (18%)

Table 3: Moderated mediation: conditional indirect associations (cluster-bootstrap CIs).

Moderator	Indirect association	95% CI
Small class ($N \approx 19$)	0.0078	[-0.0050, 0.0222]
Median class ($N \approx 29$)	0.0098	[-0.0028, 0.0255]
Large class ($N \approx 60$)	0.0159	[-0.0021, 0.0363]
Native-English	0.0160	[0.0026, 0.0330]
Non-native	0.0470	[-0.103, 0.267]

Placebos. Beauty does not predict course credits (p=0.30) or multiple-instructor status (p=0.30).

5 Threats to Identification

Unobserved confounding (charisma/teaching quality). A key threat is an unobserved factor U (charisma, pedagogical effectiveness) that may (i) increase perceived beauty via halo effects, (ii) raise participation by motivating students to submit evaluations, and (iii) increase scores directly. Such U would bias α_1 (beauty→participation) and ϕ (participation→score), inflating the product $\alpha_1\phi$. We therefore present mediated quantities as *associations*.

Sensitivity analyses. We complement cluster-robust inference with robustness checks (latent beauty, trimming, control sets) and provide code for formal mediation sensitivity (e.g., Imai et al.'s `medsens`) and Oster-style bounds in the appendix to assess how strong an unobserved confounder would need to be to attenuate the product toward zero.

Functional form and measurement. Participation is bounded; we therefore add a fractional-logit mediator. Scores are bounded; we use linear models for transparency and accompany with robustness.

Interpretation. We emphasize descriptive, decision-relevant interpretation rather than strict causality.

6 Policy Implications

Bias can arise not just from how students rate, but who rates. Larger classes and native-English instructors show stronger participation-linked associations. Institutions using participation thresholds or weights risk amplifying appearance bias. Remedies: (i) participation-adjusted reporting (residualizing predicted participation), and (ii) course-level, not instructor-contingent, incentives. In modern online systems with automated reminders, uniform nudges may attenuate—but not eliminate—composition effects when response remains well below 100%.

7 Discussion

Behavioral mechanisms. Plausible channels include halo effects (global positivity), salience (more memorable when reminders arrive), and motivated responsiveness (desire to help admired instructors). While not directly testable here, these are consistent with a participation pathway.

Generalizability. The data are from the early 2000s. Online evaluations with automated nudges likely reduce discretionary nonresponse, potentially attenuating participation mediation. Nonethe-

less, when response is incomplete and heterogeneous across courses, the mechanism remains a fairness concern.

Practical significance. The mediated component (0.017 points) is small in raw terms, but evaluation distributions often bunch near the top; differences of a few hundredths can move rankings and trigger threshold-based decisions. We therefore treat even modest participation-linked associations as policy-relevant.

Subgroup caution. For non-native instructors the conditional indirect association is imprecisely estimated (wide CI). We do not draw conclusions for this subgroup beyond noting the uncertainty.

8 Conclusion

About 17% of the beauty premium in evaluations appears to operate through participation. The mechanism strengthens with class size and for native-English instructors. This composition bias highlights risks of participation-based policies and underscores the need for adjusted reporting.

A Responsible AI Statement

This work adheres to the NeurIPS Code of Ethics. The project used publicly available, de-identified data; no new human subjects were recruited, and no PII was processed. Large language models (LLMs) were used extensively: one LLM explored questions and analyses; a second produced the draft; a third critiqued; and an LLM refined the text. To ensure safe deployment of the "AI scientist": (i) human oversight constrained analysis to variables present in the dataset and enforced a descriptive (non-causal) framing; (ii) we report assumptions, limitations, and potential harms (e.g., over-interpreting observational mediation, fairness risks of participation thresholds); (iii) we provide a clear reproducibility pathway and encourage external auditing; (iv) we avoid prescriptive personnel decisions and instead propose mitigation strategies (participation-adjusted reporting, course-level incentives). No dual-use, surveillance, or safety-critical applications are involved; environmental impact is negligible given the lightweight analyses. All code and configurations required to reproduce results are made available in the supplement.

B Reproducibility Statement

We provide all model formulas, variable definitions, and table shells in the paper. The dataset (UT Austin teaching evaluations) is public; preprocessing steps and analysis scripts (including cluster bootstrap and moderated mediation code) are released in the supplement with instructions to reproduce Tables 1, 2, and 3. Results are generated on a standard laptop in minutes (150 bootstrap resamples in the draft; we recommend $\geq$1000 for the camera-ready). Random seeds and software versions are documented.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Answer: blue**[D]**
   Explanation: Research questions and hypotheses were generated by an LLM.
2. **Experimental design and implementation**: Answer: blue**[D]**
   Explanation: Mediation specifications, robustness checks, and inference procedures were proposed by an LLM and executed accordingly.
3. **Analysis of data and interpretation of results**: Answer: blue**[D]**
   Explanation: Statistical analyses and initial interpretations were produced by an LLM; humans enforced descriptive framing and checked for overclaiming.
4. **Writing**: Answer: blue**[D]**
   Explanation: The manuscript was drafted and revised by LLMs; humans facilitated prompts and emphasis.

5. **Observed AI Limitations**: LLMs can overstate causal claims and suggest infeasible analyses; human oversight constrained scope to dataset-available variables, emphasized assumptions, and added sensitivity/robustness caveats.

Agents4Science Paper Checklist

1. **Claims**: blue[Yes] Justification: Claims (participation-linked association; moderated patterns) match Abstract/Results (Tables 1–3).
2. **Limitations**: blue[Yes] Justification: Expanded in *Threats to Identification*.
3. **Theory assumptions and proofs**: gray[NA] Justification: Empirical study; no theorems.
4. **Experimental result reproducibility**: blue[Yes] Justification: Models, moderators, bootstrap details in *Empirical Strategy* and *Reproducibility Statement*.
5. **Open access to data and code**: blue[Yes] Justification: Public dataset; code/scripts in supplement.
6. **Experimental setting/details**: blue[Yes] Justification: Variables, controls, clustering, and bootstrap documented.
7. **Experiment statistical significance**: blue[Yes] Justification: CIs and bootstrap intervals reported in all tables.
8. **Experiments compute resources**: blue[Yes] Justification: CPU-only; minutes on a laptop; details in *Reproducibility Statement*.
9. **Code of ethics**: blue[Yes] Justification: See *Responsible AI Statement*.
10. **Broader impacts**: blue[Yes] Justification: *Policy Implications* discusses fairness risks and mitigations.