Discerning Causes of Ratings Bias: A Platform for Bias Experimentation in Ratings-Based Reputation Systems

Mehul Maheshwari [1], Jacob Thebault-Spieker [2]

[1] Jacobs School of Engineering, University of California, San Diego
[2] Information School, University of Wisconsin – Madison
mmaheshwari@ucsd.edu, jacob.thebaultspieker@wisc.edu

August 2024

Abstract

As the gig and sharing economy proliferates, issues of demographic disadvantages in ratings-based reputation systems become critically important. After all, if women or non-white workers receive unfairly biased ratings, they will be disadvantaged in these kinds of marketplaces. Some prior observational work identifies these biases in the real world, whereas other work does not replicate these biases experimentally. To help characterize how and why these biases play out, we introduce a platform for streamlining the process of creating experiments in the space of five-star ratings. We aim to use this platform as a stepping stone to investigate which analytical dimensions cause and mitigate ratings bias.

1 Introduction

As of 2024, more than one-third of the US workforce, approximately 57 million workers, participate in gig work, and this number is projected to exceed 87 million by 2027 [1, 2]. Gig work offers unprecedented autonomy, allowing workers to choose when, where, and how they work [3]. The digitization of gig work has further enhanced its convenience, as over 70% of gig work is now online [2].

With the increase in digital gig work comes the rise of five-star reputation systems. These rating interfaces that are in prolific use for reviewing products to purchase or consume also help establish trust between a gig worker and requester. Unfortunately, prior work suggests that five-star ratings come with risk. When humans evaluate each other, evidence shows that demographic biases (along dimensions such as race and gender) are commonplace. Coupled with growing concerns about discrimination in gig economies, the worry becomes that five-star reputation systems become a tool to facilitate this bias [4–6, 10, 11].

The existence of these kinds of biases in real-world settings suggests that they manifest in controlled experimental settings as well [10]. However, this remains unclear. [7] found statistical confidence in the *absence* of such bias. [9] observed ratings bias with race but not with gender, yet found linguistic review bias about race and gender on Fiverr but not on TaskRabbit. [8], in an experiment of ratings received by different demographic groups of passengers in Uber, found no differences by gender. While across these three published studies results are somewhat similar, we do not know the extent or reach of experimental settings where ratings bias *is not shown*, as null results are not typically published and thus are not visible to the scholarly community.

A natural question, given these conflicting findings, is *in which settings do ratings-based reputation systems exhibit bias, or not?* Addressing this question would typically be the purview of meta-analysis, a method of identifying statistical trends across many experiments, but meta-analysis draws on larger bodies of work of scholarly work to be effective [12, 13]. In other words, a scholarly knowledge gap exists, and bridging it is essential to addressing and mitigating demographic bias in ratings-based reputation systems.

The system we demonstrate here aims to address this gap by enabling researchers to quickly run many different experiments on specific potential causes of demographic bias. By doing so, we aim to facilitate the creation of a broader body of experimental findings about settings where bias occurs, and does not, in order to define causal mechanisms of bias and bias mitigation.

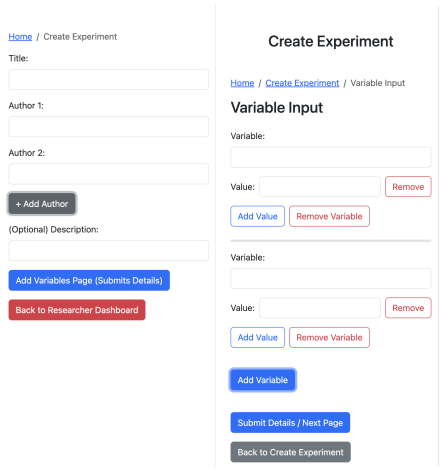

Figure 1: Create Experiment and Variables

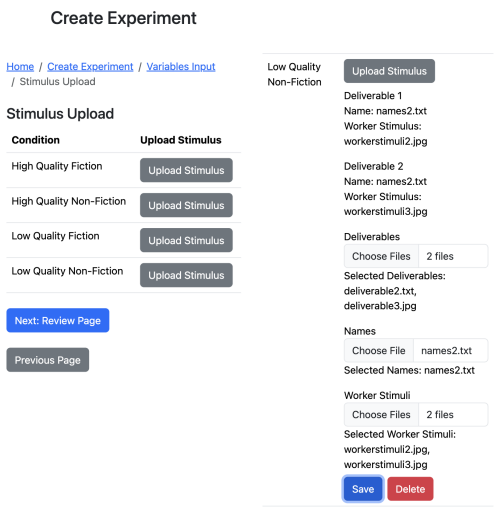

Figure 2: Stimulus Upload

2 The System

2.1 Design Rationale for the System

A key component of making this vision successful is scale — enabling many researchers to conduct many different experiments, across many different dimensions. While the prior work discussed above does not paint a clear picture of demographic bias in ratings-based reputation systems, one main takeaway is clear: demographic bias *does* occur in some settings, and *does not* occur in other settings. Which settings do or do not elicit bias, however, is uncertain. In other words, demographic bias could occur due to task specifics, the design of the five-star rating interface, the specific identities of the gig worker, or the gig consumer, or a variety of other factors as well. As such, the central design rationale for our system is to enable substantial variability in how experimental design can play out, allowing researchers the flexibility to study whatever specific dimensions of this problem they choose.

Of course, the design of the experiment is only one side of an effective study; participants need to be able to quickly join and complete the study as well. Our system is designed to rapidly deploy to either Amazon's Mechanical Turk or the web via a URL. In designing the system to both meet the experimental design goals of researchers, and quickly deploy to study participants, the intention is to connect several experimental design decisions directly to participant conditions.

2.2 Researcher Workflow

Based on this design rationale, we split the process of creating an experiment apart into its most basic components resulting in a four-page workflow for researchers.

Creating an Experiment 1

The first page focuses on experimental details, enabling a researcher to differentiate experiments from one another, and include more than one contributor.

Specifying Study Variables 1

The next page of this system asks the researcher to specify the experimental variables and values of those variables. Our system is designed to accommodate as many variables of interest as a researcher might choose, given the open-ended nature of the process.

Uploading Experimental Condition Stimuli 2

Afterward, the researcher must upload stimuli tailored to each experimental condition — or permutations of the variables specified in the previous step. For example, if the variables are Book Type (Non-Fiction/Fiction) and Quality (High/Low), a potential condition for the experiment could be Non-Fiction, High Quality.

At its foundation, conditions for five-star rating experiments need the following stimuli: a deliverable[1], a worker reference[2], and a simulated worker name. Following our design principle, our system can handle multiple deliverables per condition, as well as

[1] The "output" of a simulated gig work task, such as a written text, graphic design image, or video of a delivery worker

[2] A reference or identifier for the worker involved in the experiment, often used to track responses or attributes.

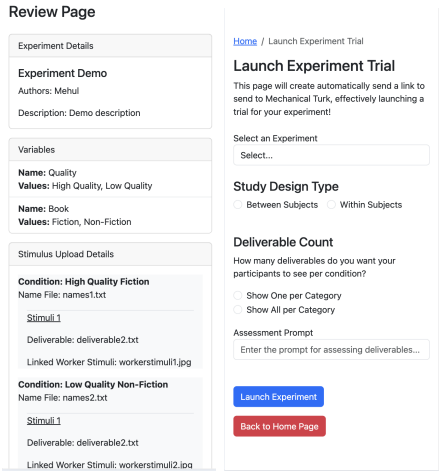

Figure 3: Review and Launch Page

any range of file types. Each deliverable must correspond to a worker reference, which can also come as any file type. We also allow researchers to randomize simulated worker names by submitting a comma-separated text file, rather than limiting researchers to a single text box.

Final Review / Submission 3

In the last step of the experiment creation process, a review page is shown to the researcher which summarizes all of the information they submitted in the previous steps. After confirming that the experimental design is correct, the researcher can save the experiment and move on to launch it.

Launching an Experiment 3

Recalling our research goals, we want to be able to rapidly configure many experiments with small differences to derive how they affect biases. To allow for this, our launch screen allows researchers to choose their experiment, and configure different tweaks. Currently, we allow for varying the study design (between/within subjects), the number of variables shown to the participant, and the assessment prompt/context. Once specified, we launch the experiment as a trial.

The process of conducting a trial starts with a HIT[3] request to Amazon's crowdsourcing platform, MTurk. Crowdsourcing platforms come with the benefit of reaching a wide, diverse audience ensuring results for these trials. Furthermore, due to the sheer quantity of workers, researchers can expect to get results for a trial quickly.

[3]Human Intelligence Task, a unit of work on crowdsourcing platforms like MTurk.

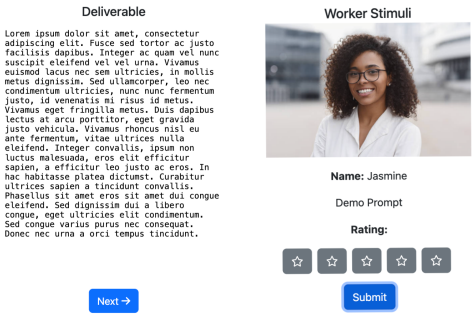

Figure 4: Deliverable and Ratings Page

2.3 Participant Workflow 4

The HIT request consists of a URL that links back to our platform. A participant gets assigned a worker ID from MTurk, which we use to store responses. Once a participant has clicked the link, they will be shown the deliverable(s) as specified in the trial. As mentioned earlier, the system is capable of handling any file type as the deliverable. In the example, we have a .txt file as the deliverable, but other valid examples could be a link to a video, or an image.

Following the deliverable, we ask participants to provide a five-star rating of the worker and deliverable. On this screen, a simulated worker is presented to the user along with a randomly selected name paired with a worker reference, and a prompt asking the participant to rate the deliverable on a five-star scale. Once the participant clicks submit, their answers are recorded. This process repeats for however many deliverables and conditions the trial specifies.

3 Discussion / Future Work

The 2017 [7] study which started our discussion enumerated several possible next studies that would help begin to understand the experimental space that elicits (or not) demographic biases in five-star rating scales. Our platform here enables the exploration of these studies, as well as a number of others. While we currently allow researchers to vary experimentation according to the variables they include, the simulated deliverable of a gig work task, and the simulated gig worker themselves, we see future work taking two primary directions. First, we intend to begin conducting experiments to help further characterize the experimental settings and situations that create bias.

Second, we intend to ultimately launch this platform more publicly, enabling other researchers to leverage this tool for their own experimentation as

well. A key aspect of that more public launch will also likely involve some additional functionality of the tool, including perhaps a mechanism to vary the rating interface more directly (e.g. asking more than a single question, using thumbs-up or thumbs-down instead of a five-star scale). We also see opportunities to help researchers more quickly understand the outcomes of their experiments, through quick analysis of whether the results show bias, or other statistical approaches to facilitating this broader goal.

Longer term, as noted above, we are also excited about conducting the eventual meta-analysis this platform will enable, though that goal hinges on establishing a broader set of experimental results.

4 Acknowledgments

The work presented here was generously supported by the Institute for Diversity Science, at the University of Wisconsin – Madison.

References

[1] Adrian, M. (2024, February 8). 25 Gig Economy Statistics, Facts, & Trends For 2024. Www.positiveaccountant.com. Retrieved from https://www.positiveaccountant.com/gig-economy-statistics

[2] Beckman, J. (2024, March 30). 2024 Gig Economy Statistics: Unveiling 85+ Remarkable Insights. The Tech Report. Retrieved from https://techreport.com/statistics/business-workplace/gig-economy-statistics/

[3] Duszyński, M. (2020, October 10). Gig Economy: Definition, Statistics & Trends [2020 Update]. Zety. Retrieved from https://zety.com/blog/gig-economy-statistics

[4] Yale School of Management. (2023). Ratings Systems Amplify Racial Bias on Gig-Economy Platforms. Retrieved from https://insights.som.yale.edu/insights/ratings-systems-amplify-racial-bias-on-gig-economy-platforms

[5] Botelho, T. L., Sudhir, K., & Teng, F. (2023). Research Investigating Bias in the Gig Economy Honored at INFORMS Conference. Yale School of Management. Retrieved from https://som.yale.edu/news/2023/10/yale-som-research-investigating-bias-in-the-gig-economy-honored-at-informs-conference

[6] Breinlinger, J., Hagiu, A., & Wright, J. (2019, July 2). The Problems with 5-Star Rating Systems, and How to Fix Them. Harvard Business Review. Retrieved from https://hbr.org/2019/07/the-problems-with-5-star-rating-systems-and-how-to-fix-them

[7] Thebault-Spieker, J., Kluver, D., Klein, M. A., Halfaker, A., Hecht, B., Terveen, L., & Konstan, J. A. (2017). Simulation Experiments on (the Absence of) Ratings Bias in Reputation Systems. Proceedings of the ACM on Human-Computer Interaction, 1(CSCW), 1–25. https://doi.org/10.1145/3134736

[8] Ge, Y., Knittel, C. R., MacKenzie, D., & Zoepf, S. (2016). Racial and Gender Discrimination in Transportation Network Companies. National Bureau of Economic Research. Retrieved November 15, 2016 from http://www.nber.org/papers/w22776

[9] Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., & Wilson, C. (2017). Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. Retrieved December 14, 2016 from http://claudiawagner.info/publications/cscw_bias_olm.pdf

[10] Bigoness, W. J. (1976). Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. Journal of Applied Psychology, 61(1), 80–84. https://doi.org/10.1037/0021-9010.61.1.80

[11] Filippas, A., Horton, J. J., & Golden, J. M. (2022). Reputation Inflation. Marketing Science. https://doi.org/10.1287/mksc.2022.1350

[12] Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to Meta-Analysis. John Wiley & Sons.

[13] Israel, H., & Richter, R. R. (2011). A guide to understanding meta-analysis. Journal of Orthopaedic & Sports Physical Therapy, 41(7), 496-504. doi:10.2519/jospt.2011.3337

[14] Jahanbakhsh, F., Cranshaw, J., Counts, S., Lasecki, W. S., & Inkpen, K. (2020). An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality and Gender. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3313831.3376860