Rethinking Measurement: Uncovering Biases in Labor Market Gaps

Adetoye Adebayo

Tufts University | Laidlaw Scholars

August 2025

Abstract

This paper reflects on my summer research experience as a Laidlaw Scholar, during which I participated in a field experiment investigating how **transparency**, specifically informing participants that they are part of a study, affects **gender bias in hiring**. In this two-sided audit study, professional recruiters on a freelance platform were randomly assigned to either **be unaware that** they were participating in a study (thinking they were completing a real hiring task) or be **explicitly informed that** the task was part of academic research. All recruiters evaluated identical sets of job applications that varied mainly by candidate gender. Our findings show that when recruiters knew they were in an experiment, **male candidates fared worse**, receiving fewer callbacks and lower recommended salaries, while outcomes for **female candidates** stayed about the same. In other words, being transparent about the study **reduced the apparent gender bias** (it made the recruiters treat men and women more equally). This suggests that if researchers always obtain informed consent in hiring discrimination studies, they might **underestimate true bias**. I describe the motivation for studying hiring bias, the methodological challenges of audit studies, our experimental design on a freelancing platform (including my role in the project), key results, ethical considerations, and personal lessons learned.

Introduction

Why does bias in hiring matter? Decades of research have documented persistent disparities in employment outcomes by gender, race, and other characteristics. For example, women and minority groups often face lower callback rates for job interviews and wage gaps compared to equally qualified male or majority counterparts. These disparities have real consequences: they can limit career opportunities and perpetuate inequality. Understanding **the causes of hiring bias** is crucial for developing strategies to make hiring fairer.

A common way researchers measure discrimination in hiring is through **correspondence audit studies**. In a classic audit study, the researcher sends out fictitious but realistic job applications (resumes/CVs) to real job postings, varying only the characteristic of interest (such as the applicant's gender or race), and then measures differences in employer responses (Pager). This method has been widely used: over 300 audit experiments were conducted between 2005 and 2020, sending nearly a million applications to test for bias (Pager 104). One famous example is a study by Bertrand and Mullainathan (2004), who mailed identical resumes under different names, some with traditionally white-sounding names and others with Black-sounding names. The result was striking: the resumes with white names received about **50% more callbacks** for interviews than those with African-American names, even though the applications were otherwise identical. This landmark finding confirmed that **racial discrimination** was (and sadly, remains) a prominent feature of the labor market. Similar audit studies have shown gender biases as well, such as women facing penalties in certain male-dominated jobs and men facing bias in some female-dominated roles (Riach and Rich F480).

Audit studies are powerful because they capture **real employer behavior** rather than just attitudes or intentions. However, they raise a critical question: **What if the employers (or subjects) knew they were part of a study?** Would they act differently? This question is at the heart of my summer project. Traditional audit studies are typically **covert**; employers are not informed they're in an experiment, to ensure natural behavior. This covert approach increases realism but means deceiving participants and not obtaining their informed consent. Ethically, that's problematic: people may object to being unwitting subjects, and indeed some **Institutional Review Boards (IRBs)** and policymakers have criticized such deception (Heckman 101). On the other hand, if you **do** tell people they're in a study, they might change their behavior because of that knowledge, which could **bias the results**. A British economist, John List, highlighted this dilemma, noting that participants who know they're in an experiment might react to that knowledge and skew the findings (List 2).

There are several reasons people may behave differently under observation. They might try to look good, especially on sensitive issues like bias. For instance, a hiring manager who knows researchers are watching might be extra careful to appear fair and non-discriminatory. Relatedly, they could be influenced by **experimenter demand effects**, attempting to guess the study's purpose and deliver the "right" answer that they believe the researchers expect. This is sometimes called the **Hawthorne effect**, referring to the idea that people change their behavior simply because they know they are being observed (Roethlisberger and Dickson). In the context of discrimination research, these effects could be especially seen: no one wants to be caught acting biased or sexist when they know they're under a microscope. As a result, fully transparent studies (where subjects know they're in a study) might show **less bias,** not because people have genuinely unbiased attitudes, but because they are adjusting their actions to avoid looking prejudiced.

To date, however, very few studies have directly compared the two approaches (transparent vs. covert) in the **same setting**. **My summer research project aimed to fill this gap.** Working with Professor Laura Gee at Tufts University, Professor Bo Cowgill from Columbia University, and Amanda Agan from Cornell University, I helped conduct a **field experiment on a freelance hiring platform** to explicitly test: *Does telling recruiters that they're in a study change how they treat male vs. female job applicants?* This meant essentially recreating a hiring audit study in two versions, one where the recruiters were **not told** it was an experiment (mimicking the traditional approach) and one where recruiters **were told** from the start that it was part of academic research. By comparing outcomes in these two conditions, we could see how much transparency might be hiding or revealing true gender discrimination in hiring.

Methods & My Role

Our study was designed as a **"two-sided audit" field experiment** on a popular online freelancing marketplace. The term "two-sided audit" means we were able to randomize aspects of both sides of a hiring interaction: the **job candidates** and the **recruiters (hiring personnel)**. Typically, audit studies randomize features of the applicants only (like names on resumes), but in our case, we also randomized whether the recruiters knew about the study.

Here's how it worked in practice. We (the research team) created two **client accounts** to simulate two different employers looking for help with hiring:

- **Professor account** – presented as a university-affiliated project, and **fully transparent** about being part of a research study. Recruiters invited to this job were told upfront that *"this is an*

academic study" and even completed an informed consent form (standard IRB consent) before doing the task. The account had a university-like branding, signaling that the work was hypothetical and for research purposes.

- **Construction account** – presented as a private company (e.g., a construction firm's HR department) with **no indication of being a study**. Recruiters invited under this account believed they were doing a real hiring task for a company. In truth, this was the "covert" condition; they were not told anything about the research or that the candidates were fake. This mimicked a normal audit where subjects think the scenario is real.

Aside from the transparency difference, the actual **task given to recruiters was identical** in both conditions. In each case, the recruiter's job was to act as a hiring evaluator: they were told that the company (real or fictional) was hiring a project manager at a mid-sized construction company, and they were given a set of **12 job applications** to review (each application included a candidate's resume and some standardized information). These applications were fictitious but carefully crafted to look authentic, based on real resumes with details adjusted for our experimental needs. The crucial manipulation in the applications was the **gender of the applicant**: we randomly assigned some resumes a male name and others a female name. Each recruiter saw a mix of male and female candidates, and across the whole sample of recruiters, each resume was used in both male and female versions to avoid any resume-specific effects. The recruiters were asked to provide feedback on each candidate, including:

- Whether they would **recommend interviewing** that candidate (this is essentially a "callback" decision, yes or no).

- A suggested **salary offer** if the company were to hire that person (a specific annual salary figure).

- A **willingness-to-pay (WTP)** assessment, meaning the maximum salary the recruiter thinks the company should be willing to pay for that candidate. In other words, "how high would you go to hire this person if needed?".

- Additionally, they could leave comments on each candidate and were prompted to rate how confident they were.

To make the task realistic, we even included aspects like some resumes showing a current salary from the candidate's last job (a detail varied as part of a related research question on salary history, beyond the scope of this reflection). Importantly, **recruiters in both conditions were paid** for their time, at

an hourly rate they agreed to (their normal market rate), plus a performance bonus for completing the evaluations. This was done to ensure we weren't exploiting their time unethically. Even in the "not told" group, they got paid just as they would for any real contract work.

My role in this project was primarily on the **"Professor account" team**. As a Laidlaw Scholar, I was essentially acting as a research assistant and project coordinator under the guidance of the professors. **Using a freelancing platform to recruit freelancers** (in our case, professional recruiters) was a novel experience for me. It required a blend of technical, organizational, and interpersonal skills, and it ended up being quite an adventure.

Recruiting the recruiters: We started by identifying a large pool of potential recruiter participants. We searched profiles with keywords like "recruiter," "human resources," "staffing," etc., to find people with relevant experience. From this, the team compiled a list of thousands of freelancers worldwide who could fit our criteria (e.g., having some hiring experience, being based in or familiar with the U.S., and charging a feasible hourly rate). We then set up our two accounts to begin sending out **invitations** to these freelancers to apply to our job posting. The invitation explained the gist of the task. In the **not-told (Construction) condition**, the invite message was phrased as if from a company: "We're hiring a project manager... we have 12 applications for you to review and give feedback..." etc., with no mention of research. In the **told (Professor) condition**, the invite message included language that this was part of a research project at a university and that their feedback would help an academic study, and it asked them to sign a consent form approved by an IRB. Aside from that, both messages described the same task and offered the same pay structure.

Schedule and scale: One of my big responsibilities was helping manage the **mass invitation process**. We needed a lot of recruiters to get a sufficient sample size, because not everyone invited would accept and complete the task. Over the first two weeks of July (July 1–14, 2025), we sent out **hundreds of invites each day**, ramping up to over a thousand per day at times. We had an internal goal of reaching roughly 16,000 invitations in total to yield the desired number of participants. In practice, on Day 1 (July 1), we invited 1,096 freelancers, on Day 2, we sent 1,232 invites, and we hovered around ~1,100 per day for most of that week. We then slightly increased the volume in the second week – for example, on July 8, we sent about 1,204 invites, and on July 9, about 1,221. By the end of July 14, we had sent **16,197 invitations** in total. (Table 1 below presents a summary of the number of invitations sent each day during the main recruitment period.)

Table 1. Daily Invitations Sent (July 1–14, 2025)

Date	Invitations Sent	Cumulative Total
Tue July 1	1,096	1,096
Wed July 2	1,232	2,328
Thu July 3	1,096	3,424
Fri July 4	1,096	4,520
Sat July 5	1,096	5,616
Sun July 6	1,096	6,712
Mon July 7	1,096	7,808
Tue July 8	1,204	9,012
Wed July 9	1,221	10,233
Thu July 10	1,159	11,392
Fri July 11	1,183	12,575
Sat July 12	1,207	13,782
Sun July 13	1,199	14,981
Mon July 14	1,216	16,197

Team coordination: Our invite schedule was carefully planned, not just in volume but also in *who* was doing *what*. We had a team of RAs (including myself) and occasionally the professors themselves logging in to send invites or monitor responses. Each day, a few team members were assigned to handle a batch of the invites, as well as to keep an eye on the **inboxes** of the accounts for any incoming messages from freelancers. For instance, if an invited freelancer wrote back with a question (“Can we discuss details via a call?” or “Is this a one-time project or ongoing?” etc.), we needed to reply promptly with a polite and scripted answer. Since I was on the Professor account team, I handled these communications. We had **pre-written script responses** for common questions and scenarios to ensure consistency and avoid saying anything that could bias the subject or reveal more than intended.

Examples of these scripts included responses to: “Why only 1 hour contract?” (we’d reassure them that it’s a short task, but we’re still interested), “Can I have more pay?” (sadly, we had to stick to the offered rate), etc. I found myself using these canned responses a lot; it felt a bit robotic at times, but it taught me how a large experiment requires a standardized approach to communication. It also improved my ability to write professional, clear messages quickly.

Every day, I would log in around the scheduled times (we often aimed to send invites at set intervals like 9 AM, 12 PM, 3 PM, etc., to spread them out) and coordinate with my teammates via a group chat. We tracked which invitation “batch numbers” we were sending, referring to internal IDs for blocks of resumes/invites.

After the initial invitation wave, we also did a **reminder wave**. For those recruiters who received an invite but hadn’t responded after about a week, we sent a follow-up message saying, essentially, “We’re still interested if you are – you can still accept the offer to work on this task.” According to our plan, starting around July 15, we began these reminders. In the first week of reminders, only the Professor account sent reminders (due to some minor issues on the Construction account). By July 21, we also sent reminders on the Construction account side. The reminders helped bump up our recruitment a bit, as some people who missed or ignored the first invite came on board when prompted again.

Executing the experiment: Once a recruiter accepted the contract, we would send them the materials (the set of resumes and the evaluation form link).

Overall, my role was a mix of **operational management and support:** I was simultaneously an “HR coordinator” dealing with the freelancers and a “data steward” ensuring the experiment ran according to protocol. I felt like we were running a high-stakes operation, and I was right in the middle of it. One of the most gratifying moments for me was when we reached our target number of completed evaluations, seeing the dataset come together after weeks of effort. By the end, we had 76 recruiters in the not-told (Construction) condition and 56 in the told (Professor) condition who completed the task, each having evaluated 12 applications (so data on 1320 candidate evaluations in total). The imbalance (76 vs 56) was because we intentionally invited more people in the first group, anticipating that some might drop out once they realized it was a research study in the second group, an expectation that did bear out to a degree (fewer recruiters accepted the “academic study” invite, perhaps not wanting their first contract to be an experiment, as the professors hypothesized).

Findings

Working with the research team's preliminary analysis (which later became a working paper), I learned how our experiment answered the core question: *Did telling recruiters they were in a study change the gender-related outcomes?* The answer was a clear **yes**. Transparency had a notable effect, specifically on how **male candidates** were treated.

Callbacks (Interview Rates):

When recruiters thought it was real, men had a small edge in callbacks. Recruiters were slightly more likely to recommend interviewing men over women. But when recruiters knew they were being studied, this advantage disappeared. In fact, men's callback rates dropped a bit, while women's stayed the same. This leveled the playing field, and in some cases, women even edged out men.

Salary Offers (Willingness to Pay):

Pay recommendations showed the clearest shift. In the real condition, men were offered higher salaries than women, about a 6% difference, or roughly $100k vs $94k (Agan, Cowgill, and Gee). In the transparent condition, male salary offers dropped by about $4,000 on average, while women's stayed the same. This meant the gender pay gap basically closed, rising from about 94 cents on the dollar to 98 cents.

Effort and Motivation:

Importantly, recruiters in both groups put in similar effort. They left notes, checked outside resources, and engaged with the task. In fact, those told it was a study sometimes left more comments. So the behavior change wasn't because they "didn't care", it was a real shift in how they evaluated candidates.

Other Observations:

Some recruiters were less likely to accept invitations when told it was research, especially newer users. But since we randomized invitations, the groups ended up being balanced in experience and background.

What It Means:

The main takeaway is that transparency reduced the appearance of gender bias. Men lost their small advantage when recruiters were aware of being observed, while women's outcomes stayed steady. This shows a trade-off in research design: transparency is more ethical, but it may hide or understate real-world inequalities (Agan, Cowgill, and Gee). If we only studied bias in transparent settings, we might wrongly conclude that gender inequality in hiring is minimal. Our study highlights how simply being observed can change people's choices, especially in sensitive areas like gender and pay.

Ethical Considerations

Working on this project made me think a lot about research ethics. Our main challenge was the trade-off between being honest with participants and keeping the study realistic. If we told recruiters everything up front, their behavior might change, but not telling them raises concerns about consent.

Informed Consent vs. Realism:

In the "Professor account" condition, recruiters signed a consent form and knew they were in a study. This was ethical, but it risked changing how they acted. In the "Construction account" condition, recruiters didn't know until the end. That gave us more natural results, but it also meant we withheld information. To reduce harm, we paid all recruiters fairly, didn't give false feedback, and fully debriefed them afterward.

Why It Was Justified:

We believed this approach was acceptable because discrimination in hiring is an important issue that can't be studied well if participants know too much. Our work tested not only hiring bias but also how transparency itself affects results. The IRB approved this balance, seeing that the benefits of uncovering hidden bias outweighed the limited deception.

Protecting Participants:

We kept tasks reasonable (about an hour, flexible deadlines), and recruiters could drop out anytime. In the transparent group, this meant leaving the study; in the hidden group, it looked like ending a contract. Either way, they had the choice. Everyone was paid, no sensitive data was collected, and no one's job opportunities were affected since we invited individuals directly.

Debriefing:

In the end, we explained the study to all participants, especially those who didn't know. Most were curious and even glad to have contributed to research on discrimination. This step helped restore trust and avoided any lasting harm.

Takeaway:

The biggest lesson I learned is that research like this is a balancing act. Full transparency protects ethics but may hide bias; full deception shows real behavior but risks trust. Our study tried both and showed just how much transparency can change outcomes. For me, this highlighted the tension between ethics and truth in research, and how important it is to handle that tension with care.

Personal Reflection

Beyond the data and the research questions, this summer's project was transformative for me on a personal level. I came in as an eager undergraduate, but I left with a far deeper sense of what it means to be part of a research team and to contribute meaningfully to knowledge.

Teamwork and Collaboration

This was my first time working on such a large, coordinated project, and I quickly saw how important communication was. Each morning, our team updated one another on progress, problems, and next steps. I started out hesitant to speak up, but I later on realised that we all want the best for each other.

Discipline and Protocols

At first, following strict scripts for every message felt stifling, but I came to appreciate the consistency that science demands. Carefully documenting every step, keeping logs, and respecting the protocol showed me how structure ensures credibility. That shift, from "winging it" to systematic precision, has reshaped how I think about research.

Problem-Solving and Adaptability

No project goes perfectly. When invites expired unexpectedly or freelancers made mistakes, I learned to stay calm, improvise solutions, and help build contingency plans. Each challenge pushed me to think more like a problem-solver than just a student following directions.

Leadership and Initiative

Over time, I began to step forward more naturally, sending reminders, actively responding to the inbox, and even helping new team members in areas where they were confused and were a bit shy to ask the professors. Managing the "Professor account" process became my responsibility, and I realized that leadership is as much about noticing needs and stepping up as it is about formal titles.

Meaning and Resilience

Most inspiring of all was the project's subject matter: bias and fairness in hiring. Seeing how discrimination operates in real processes made the work feel urgent and socially important. At the same time, the grind, sending hundreds of invites, staying organized, taught me resilience. I now understand that research is equal parts discovery and persistence.

Closing Reflection

In sum, this summer gave me not only technical and professional skills but also the confidence that I can contribute meaningfully as an undergraduate researcher. I discovered that I thrive on problem-solving, I enjoy structured teamwork, and I am motivated when research connects to

real-world impact. These lessons will stay with me long after the summer, and they have only deepened my commitment to pursuing research as a way to engage with pressing social issues.

Conclusion

In conclusion, this summer project underscored a central insight: **the act of measurement can influence the phenomenon being measured**, telling people they're being watched changes what you see. This paradox is important for social scientists to grapple with. Our study provides evidence that in hiring discrimination research, **more honesty (with participants) can lead to less honesty (in their behavior)**, a finding that will inform how future experiments are designed. For me, as a student researcher, the project was transformative. I not only learned about the topic at hand (gender bias and experimental methods) but also about the process of inquiry and my own capabilities. I end this paper with a sense of fulfillment and excitement. The experience has been an honest and insightful reflection on how research is done and why it matters. I look forward to taking these lessons into my future studies and continuing to explore ways to uncover the truth.

Bibliography

1. Agan, Amanda, Bo Cowgill, and Laura Gee. *Transparency and Bias: Evidence from a Two-Sided Field Experiment on Hiring.* Working Paper, 2025. SSRN.

2. Bertrand, Marianne, and Sendhil Mullainathan. “Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.” *American Economic Review*, vol. 94, no. 4, 2004, pp. 991–1013.

3. Heckman, James J. “Detecting Discrimination.” *Journal of Economic Perspectives*, vol. 12, no. 2, 1998, pp. 101–116.

4. List, John A. “The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions.” *Journal of Political Economy*, vol. 114, no. 1, 2006, pp. 1–37.

5. Pager, Devah. “The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future.” *Annals of the American Academy of Political and Social Science*, vol. 609, no. 1, 2007, pp. 104–133.

6. Riach, Peter A., and Judith Rich. “Field Experiments of Discrimination in the Market Place.” *Economic Journal*, vol. 112, no. 483, 2002, pp. F480–F518.

7. Roethlisberger, Fritz J., and William J. Dickson. *Management and the Worker.* Harvard University Press, 1939. (Origin of the “Hawthorne effect.”)