# Incentivized Ratings Design in the Field

Elisa Macchi[*†]

Brown University

December 19, 2025

**Abstract**

The incentivized ratings design (Kessler et al., 2019) has emerged as a transparent and incentive-compatible alternative to correspondence and audit studies for measuring discrimination and eliciting decision-makers' preferences. While originally implemented in online hiring settings in the U.S., I argue that its structure is especially well suited to contexts where decisions are made in person and where traditional audit methods are infeasible —features common in developing countries. This article reviews the expanding body of research using incentivized ratings, highlights conceptual innovations, and synthesizes lessons across 38 studies conducted between 2019 and 2025. Drawing on these applications and on my field implementations, I provide practical guidance on implementing this design in low-resource settings.

## 1 Introduction

Measuring decision makers preferences is central to research on discrimination. In many settings—particularly in low-income countries—evaluations occur informally or face-to-face

[*]Email: elisa_macchi@brown.edu

[†]I am grateful to Judd Kessler, Corinne Low, and Colin Sullivan for sharing a list of papers applying their incentivized ratings approach which formed the starting point for the literature review section of this paper. Colin Sullivan, in particular, also generously took the time to explain the IRR approach and its practical implementation in 2019. I owe special thanks to David Yang for first bringing the method to my attention. The paper draws from conversations with Livia Alfonsi and Claude Raisaro, as well as inputs from Emanuele Colonnelli, Francesco Ferlenga, Eliana La Ferrara, and Jon Roth. Adi Jahic and Ananya Narayanan provided outstanding research assistance.

rather than through written or online applications. Employers hire workers who show up at their premises, loan officers interview applicants who arrive without documents, and credit, hiring, or admissions decisions rely heavily on referrals, social networks, and brief in-person interactions. These features make it difficult to observe early-stage screening decisions, to hold information sets constant, understand the applicants sets, to experimentally manipulate characteristics as in traditional correspondence and audit studies. As noted in a recent literature review by Heath, Bernhardt, Borker, Fitzpatrick, Keats, McKelway, Menzel, Molina and Sharma (2024), limited access to raters and the absence of formal application processes make research on discrimination particularly challenging in poor countries.

The incentivized ratings design (Kessler, Low and Sullivan, 2019), or incentivized résumé rating (IRR), offers a promising alternative. In this paradigm, decision makers evaluate openly hypothetical but realistic applicant profiles whose attributes are experimentally varied. Their evaluations directly determine which real candidates they are referred to through a matching algorithm. The key innovation of IRR design is that, because raters understand that their assessments influence the real referrals they receive, the design elicits truthful preferences without deception.

Since its introduction, researchers have applied incentivized ratings designs across a wide range of domains—including labor markets, credit, investment, academia, and dating. These studies have introduced new outcome measures, adapted the design to low-literacy environments, combined IRR with complementary interventions, and developed new strategies for constructing profiles and delivering incentive-compatible matches. This includes a growing number of developing-country contexts where formal application processes are limited. Despite this emerging body of work, there is little systematic guidance on how to implement incentivized ratings designs in the field.

This paper serves two purposes. First, it provides a conceptual and methodological overview of the incentivized ratings design, clarifying how it relates to and differs from correspondence studies, audit studies, and other experimental approaches to measuring dis-

crimination and preferences. I highlight why the design is especially suited to studying in-person decisions and why it performs well in settings characterized by informality, limited documentation, and heavy reliance on referrals. Second, drawing on 38 studies conducted between 2019 and 2025 as well as my own field experience, I offer practical guidance for researchers. I also discuss ethical considerations, experimenter demand, external validity, and suggestions for analyzing IRR data.

# 2 The Incentivized Ratings Paradigm

## 2.1 Original Design: Kessler, Low, and Sullivan (2019)

Kessler, Low, and Sullivan (2019) introduced the incentivized ratings design as a transparent and incentive-compatible method for eliciting employer preferences without deception. The central idea is straightforward: decision makers evaluate hypothetical but realistic applicant profiles whose attributes are experimentally varied, and their evaluations directly determine which real candidates they later receive through a matching algorithm. Because raters know that their assessments influence the referrals they obtain, the design elicits truthful preferences while allowing researchers to randomize candidate characteristics.[1]

The original implementation consisted of three components. First, the authors assembled a pool of real job candidates—graduating seniors from the University of Pennsylvania—whose résumé information was used to construct realistic but fictitious profiles. Second, human resources officers at participating firms evaluated these profiles on a Likert scale across various outcomes. The attributes shown on each résumé (such as gender, GPA, major, or work experience) were cross-randomized, allowing clean identification of how each trait affected evaluations. Third, a machine-learning matching algorithm used the elicited ratings to determine which real candidates would be referred to each firm. Higher-rated

[1] See Litwin and Low (2025) for a discussion on the incentive compatibility of IRR designs.

hypothetical profiles increased the likelihood of receiving a meeting with their corresponding real candidates, thus aligning the incentives of raters with truthful reporting.

## 2.2 Advantages and Limitations Relative to Alternative Methods

The incentivized ratings design offers three key advantages over traditional approaches to measure discrimination using field experiments. The main improvement is that it reduces deception (more on this later). Different from correspondence and audit studies, raters in IRR know they are evaluating hypothetical profiles. The ethical concerns associated with deception in discrimination studies have been raised in Duflo and Bertrand (2016), and are discussed thoroughly in Litwin and Low (2025). This transparency is particularly valuable in markets or settings which are small, like in developing countries, where repeated deception could damage researchers' reputations and compromise future studies.

A second advantage is the richness of the data generated. Correspondence studies typically yield a single binary outcome per application—such as whether a candidate receives a callback—limiting researchers' ability to analyze how different attributes shape evaluations. In the incentivized ratings paradigm, each rater evaluates many profiles and potentially across multiple outcomes. This repeated structure increases statistical power and allows for fine-grained analysis of how specific characteristics affect ratings, interview choices, or other elicited outcomes. The complementary measures, such as beliefs about candidate behavior or assessments of non-salient attributes, providing a multidimensional view of the decision-making process and allowing a simple approach to tease out mechanisms.

A third advantage is the high degree of control the design affords over profile content. Researchers can precisely randomize candidate attributes, hold all other information constant, and generate within-rater or within-profile comparisons. Such control is difficult to achieve in audit studies, where actor characteristics inevitably vary in subtle ways and also raise concerns about the effect of conscious or subconscious biases in non-double-blind designs, as discussed in Heckman (1998) and Duflo and Bertrand (2016)—and in correspondence studies,

which are constrained by existing résumé formats and information content. The flexibility of the incentivized ratings design enables clean implementation of mechanism tests, information treatments, and structured comparisons of alternative models of discrimination, all while preserving tight experimental control.

A fourth advantage, also highlighted in Litwin and Low (2025), is that IRR enables researchers to study preferences within subject pools that are normally difficult to access through traditional field experiments. Many important decision makers—such as HR officers, credit officers, investors, venture capitalists, or members of admissions and selection committees—may not be the individuals who screen applicants in the first stage or who respond to résumés in correspondence studies. Nevertheless, they can meaningfully participate in incentivized ratings exercises because the task mirrors decisions they routinely make and because the matching mechanism provides them with tangible value.

There are also limitations. Most notably, raters are aware that they are participating in an experiment, which may generate reputational or social concerns. Although the referral-incentive mechanism mitigates experimenter-demand effects—since truthful evaluations improve the quality of referrals—it is difficult to assess whether experimental incentives are sufficiently strong to dominate social incentives. Moreover, as discussed in Litwin and Low (2025), the design is not incentive compatible in the standard sense because respondents do not observe the algorithmic matching process. This creates scope for strategic behavior, such as manipulating stated preferences to increase the probability of an exceptional match or to hedge against low-quality matches.

While these concerns are valid, their severity depends on the context and the actual implementation of the design. In particular, they are attenuated when profile attributes are cross-randomized across identity characteristics and other relevant dimensions. Under such implementation, respondents need not distort evaluations along identity lines—for example, by expressing gender bias—to avoid low-skill matches, since identity and ability vary independently across profiles.

# 3 IRR for In-Person and Developing-Country Settings

## 3.1 Challenges of Studying Discrimination in Poor Countries

Studying discrimination in developing countries presents a distinct set of challenges compared to higher-income settings. A central obstacle is that many major economic decisions are conducted face-to-face from start to finish, whereas in richer countries initial contacts often occur through email, mail, phone, or online platforms. Because employers and other raters rarely rely on mailed or online applications, it is difficult for researchers to observe preferences or credibly infer how observable traits translate into offers.

Other obstacles compound this difficulty. Establishments, including firms and credit institutions, are typically small, vacancies are often filled through personal networks (Beaman and Magruder, 2012; Heath, 2018), and family ties or community reputation frequently substitute for formal credentials. Candidates also seldom submit formal documents. In Kampala, for instance, loan candidates at most formal or informal moneylenders queued in person and presented their case without pre-compiled paperwork (Macchi, 2023). Even in more financially developed contexts such as Indonesia or India, in-person interactions remain central (Bursztyn, Ferman, Fiorin, Kanz and Rao, 2017; Cole, Kanz and Klapper, 2015; Fisman, Paravisini and Vig, 2011). Similarly, in Kampala, blue-collar workers simply presented themselves at firm gates to ask for jobs, a feature common to spot labor markets across developing countries (Rosenzweig, 1988; Cefala, Kaur, Schofield and Shamdasani, 2024; Breza and Kaur, 2025; Macchi and Raisaro, 2025).

Unsurprisingly, the volume of research on discrimination in developing countries remains relatively small (Gentile, Kohli, Subramanian, Tirmazee and Vyborny, 2023). As Heath et al. (2024) summarized in the context of gender discrimination in labor market research, "the volume of research on demand-side barriers is smaller than on supply-side barriers in developing countries, which may be related to the higher costs of conducting well-identified studies with employers than with (potential) employees."

Duflo and Bertrand (2016), in their review of field experiments on discrimination, identify only three correspondence studies conducted outside high-income settings. One example is Banerjee, Bertrand, Datta and Mullainathan (2009) who employs a résumé audit methodology similar to that of Bertrand and Mullainathan (2004) testing whether firms in India's fast-growing service sectors— software and call centers— discriminate against equally skilled candidates from historically disadvantaged caste groups and Muslims. These sectors, however, are not representative of the jobs that dominate most labor markets in low-income countries. Another exception is Galarza and Yamada (2014), who sent fictitious CVs in response to job advertisements in a leading Lima newspaper to examine whether firms discriminate against candidates of indigenous origin in favor of those with white backgrounds. Similarly, Maurer-Fazio (2012) conduct a large-scale field experiment in China, investigating how firms respond to online job-board applications from ethnic minority candidates. Yet these studies, too, involve more structured firms and formal recruitment channels than are typical in most developing-country labor markets and, in turn, likely offer a non-representative picture of discrimination in the broader labor market.

To study less formal settings, researchers in development economics have sometimes relied on in-person audit studies or "mystery shoppers," in which trained actors pose as customers or candidates to track how they are treated (e.g., Garz, Giné, Karlan, Mazer, Sanford and Zinman, 2021). These studies, however, face similar concerns about deception as correspondence studies and come with the additional limitations discussed above.

## 3.2 Why IRR Is Particularly Well Suited in These Settings

The features that make it difficult to study discrimination in low-income and highly informal markets also make the incentivized ratings design especially appealing in these environments. First, the absence of formal application channels and the prevalence of face-to-face interactions limit the feasibility of correspondence studies, which rely on written submissions, email communication, or online platforms. IRR avoids this constraint because the

profiles are presented directly to decision makers by the research team, without requiring any existing infrastructure. In fact, the design accommodates the limited documentation and heterogeneous information environments that characterize many low-income markets because researchers construct the profiles, and they can standardize the information set across candidates/institutions.

Most notably, the incentive structure of IRR designs is particularly well suited to referral-based markets, where employers and other decision makers already rely heavily on networks and personal recommendations. The exercise can be naturally framed as an informal matching or referral service— one that closely mirrors existing market practices and enhances engagement.

Finally, avoiding deception is a particularly important advantage in small and densely networked markets where employers, loan officers, and other decision makers frequently interact with each other and with researchers.

# 4 IRR Applications in Economics

## 4.1 Scope and Growth of IRR Applications

The empirical use of incentivized ratings design has grown substantially since the original Kessler et al. (2019) article. Table 1 summarizes all studies we identified that leveraged the incentivized ratings design by year, country, and sector/topic. As of September 2025, we identify 38 studies spanning 15 countries. Of these studies, 13 are published, 19 are working papers, and 6 are pre-registered and concluded. The pace of output has accelerated: whereas only 4 studies were posted or published before 2022, there have been 34 studies since. The global distribution of these studies, along with the income levels of the countries in which they were implemented, is shown in Figure 2.

To identify the list, we proceeded as follows. We began with a list of studies that self-reported as IRR studies to the IRR platform (https://irr.wharton.upenn.edu/). We

verified these papers do in fact meet our criteria of the IRR methodology before adding them to our list. We then searched Google Scholar for papers that cited KLS (2019) and added papers that met these criteria to our list. We then repeated this process for all the well-cited papers in our list. Afterward, we searched Google Scholar using keywords such as "incentivized resume rating", "incentivized rating design", "resume matching", and the relevant shortenings and combinations of these phrases. After sufficiently exhausting these searches, we searched through the AEA Registry for other studies that (plan to) use the IRR methodology.

Some papers are exact replications of the original paradigm applied to a new sample or new contexts such as the credit market, investment markets, or academia. However, some papers implement experimental twists that range from changing the outcome measurement to more complex modifications in the design or the combination with complementary randomized interventions to test for mechanisms. These variations raise the question of what qualifies as an incentivized ratings design study. A study met our criteria for inclusion if it 1) did not use deception, 2) used hypothetical resumes, and 3) incentivized responses through some mechanism. The latter is a less objective criterion. For example, the most borderline study is Gallen and Wasserman (2023), which provides personalized advice on finding an academic mentor based on participants' responses rather than referrals.

## 4.2 Themes in IRR Applications

Given these criteria, the plurality of papers we identify study discrimination in labor markets in the U.S.. About 14% of papers are exact replications of the original paradigm (no experimental twists, through an online platform, and on labor markets), while the remaining introduce some variation in the design (N = 20), setting (N = 17), implementation (N = 21), or country of study (N = 21). A review of this literature shows that, beyond avoiding deception, the IRR design enables research that would be infeasible or impossible using traditional correspondence studies or other field experimental methods. In particular, a

non-trivial share of incentivized resume rating studies focuses on raters making decisions in person (32%) or in developing countries (34%), where correspondence studies face practical constraints.

As noted above, several studies have also applied IRR to settings that could not have been studied with traditional correspondence studies because choices are implicit or informal (i.e., there is no application procedure). Low (2024) leverages this approach to study the impact of age on women's appeal in the dating market, Chan (2022) studies discrimination against doctors, and Feng et al. (2024) examines discrimination against female startup founders.

Incentivized ratings designs also enable researchers to study introspective choices or preferences. Nielsen and Rigotti (2022) uses the design in a lab setting to study incompleteness of preferences over monetary gambles. Subjects rank gambles; these rankings are used to estimate preference structures, and payments are based on the inferred preferences. Gallen and Wasserman (2023) elicits students' preferences over mentor attributes, finding that female students are willing to trade off occupational match to access a female mentor.

As with other experimental designs, IRR can be flexibly augmented to test discrimination models—for example, to separate taste-based from statistical discrimination by varying the amount of information presented in the applications (Macchi, 2023). A key advantage of implementing this test in the IRR setting relative to a correspondence study is that researchers have control over the amount of information included in profiles, allowing them to credibly explain that some profiles will include certain information while others will not. This enables a cleaner test of statistical discrimination relative to the high-quality versus low-quality candidate approach. Bohren et al. (2025) develop an IRR-based design to measure systemic discrimination, capturing how discrimination in sequential decisions indirectly contributes to disparities (iterated audit design).

## 4.3 IRR in Developing Countries and In-Person Decisions

IRR designs have been increasingly applied to study labor market discrimination and preference elicitation in developing countries, where traditional audit studies face greater implementation challenges. Several studies have adapted the methodology to local contexts while introducing novel design features.

Macchi (2023) examines whether obesity in Uganda functions as a signal of wealth and investigates the causal effect of additional information provision on (positive) weight discrimination. Participants evaluate randomly assigned, weight-manipulated portraits serving as CV identifiers to estimate the effect of a higher BMI on credit outcomes. A 2×3 design cross-randomizes the obesity signal with three information sets (no financial information, low-quality borrower, high-quality borrower). This allows the researcher to test whether expanding the set of observable signals reduces reliance on focal discrimination attributes, providing evidence on the mechanisms of statistical discrimination.

A few studies focus on gender discrimination in hiring. Gentile et al. (2023) study barriers to entry for women and the gender gap in job search in urban Pakistan using a paired comparison method with real CVs from an online job platform. Their approach involves pairing similar candidate CVs based on education and experience, anonymizing the profiles, and then randomly reassigning names, gender, and other personal characteristics. Low et al. (2023) conduct a resume ratings experiment in India to estimate hiring discrimination by gender and other minority status. They find that explicit identity cues on resumes, such as gender, caste, and region, produce only small and statistically insignificant effects on hiring outcomes when controlling for recruiter and application characteristics. Macchi and Raisaro (2025) test for the effect of monitoring and safety frictions on demand for female workers in three male-dominated sectors in Uganda. Using an incentivized resume rating design with a meeting choice outcome and a matching mechanism similar to mentorship rather than employment, the study shows that a significant share of the unmet demand for

female workers is driven by beliefs about women's trustworthiness combined with binding monitoring constraints.

Bartŏs, Castro, Czura and Opitz (2024) analyze statistical discrimination in startup funding in Uganda. Loan officers can use real money to buy more information about entrepreneurs after reviewing their initial business pitch decks. They then select the "best" business, allocate investment amounts from their endowment, and decide which information to purchase. An interesting innovation is that the experimental incentives are structured around the startups' actual performance two years later, increasing the stakes and realism of the decision environment. By eliciting beliefs about various aspects of the application—an advantage discussed in Kessler et al. (2019)—this paper sheds light on the mechanism behind discrimination, namely beliefs over implementation constraints due to family responsibility. Other work has used incentivized rating designs to evaluate startup founder preferences over investment or partners (Colonnelli, Cruz, Pereira-Lopez, Porzio and Zhao, 2025a; Ebrahimian and Zhang, 2025). Colonnelli et al. (2024a) elicit preferences for government participation in China's venture capital market, finding that private firms dislike government-tied investors. Colonnelli, Loiacono, Muhumuza and Teso (2024b) investigate decisions in public sector procurement in Uganda.

IRR design has been used to gain insights into how to increase the participation of women in the labor market. Martin (2024) introduces a double IRR design to study gender discrimination and the impact of information constraints in Iraq. In the first stage, employers rate six job candidate profiles, which allows the authors to estimate sector-specific hiring rates for women. In the second stage, jobseekers rate six job profiles, and the treatment group randomly receives information about female hiring rates revealed in the first IRR. The paper finds that disclosing employer preferences can meaningfully improve matching by correcting women's underestimation of their employment prospects. Del Carpio and Fujiwara (2023) investigates whether using gender-neutral language in job advertisements increases female representation in tech sector job applications in Spanish-speaking Latin

America. In Colombia, Bustelo, Diaz, Lafortune, Piras, Salas and Tessada (2023) study preferences regarding job schedule flexibility (flexible scheduling and part-time employment); they employ a binary-choice conjoint design where participants view ten menus with three job options each and select their single preferred job per menu, rather than rating all options.

Abebe, Caria and Ortiz-Ospina (2021) and Ndayikeza (2025) both conduct field experiments in East Africa to investigate employer hiring preferences, focusing on different aspects of selection. Abebe et al. (2021) conduct a field experiment in Ethiopia to investigate how firms select talent from pools of real job applicants by understanding whether firms can improve the talent of their applicant pool by offering incentives to apply. They use an incentivized resume rating design to test whether managers recognize that treatment group applicants—those who received subsidies to apply—are higher quality than control applicants. Firm managers are presented with three real CVs, each from a different experimental group, and asked to rank them according to suitability for a clerical position. Ndayikeza (2025) assess employer preferences for work experience among college graduates in Burundi. This approach uses real CVs but modifies graduation dates to simulate a year of "post-graduation" experience and duplicates resumes to add or omit low-skill work during that period. Interestingly, CVs are delivered to local employers by an HR company with existing employer relationships—rather than academic researchers— enhancing realism and reducing detection concerns by employers. The results show that having any postgraduate work experience, even in a low-skilled job, significantly improved employer assessment scores.

## 4.4 In-Person Decisions

Most relevant for this paper, IRR allows researchers to study discrimination in settings with in-person decision-making. 32% of the IRR studies we identify in our literature review focus on in-person decision contexts, with this share rising to 75% in developing countries. This cross-country heterogeneity in implementation modes is shown in Figure 3.

In poor countries, IRR has been applied to study discrimination where decisions are made in person, especially in the labor market (Macchi and Raisaro, 2025; Ndayikeza, 2025; Gentile et al., 2023; Abebe et al., 2021; Archibong et al., 2025; Beam et al., 2024) and the credit market (Macchi, 2023; Bartős et al., 2024). By eliciting incentivized evaluations of hypothetical applicant profiles directly from real decision-makers, IRR replicates key features of in-person screening while maintaining experimental control. These contexts share common features that make correspondence studies infeasible: candidates present themselves in person rather than submitting written applications, the interview process is often unstructured, and the relevant raters are directly accessible to researchers through in-person interviews.

Figure 1: Published incentivized ratings design studies by year, country, and sector

# 5 Designing and Implementing an IRR Study in the Field

This section offers practical guidance for researchers planning their own incentivized ratings studies, with particular attention to in-person implementations and low-resource environments. The discussion draws on my experience implementing IRR as well as lessons and innovations from the growing literature.

The incentivized ratings design relies on two key connected inputs: (1) access to information about a pool of real candidates, and (2) a pool of raters who value referrals to those candidates. Importantly, Kessler et al. (2019) generously make their survey scripts, matching code, and implementation guides available,[2] and additional implementation advice is summarized in Litwin and Low (2025). These resources substantially lower the barrier to entry for researchers planning IRR studies.

[2] https://irr.wharton.upenn.edu/researchers/

The focus here is on adaptations that are particularly relevant for in-person and developing-country settings. Implementation in these contexts differs in important ways. Lower literacy rates or limited internet connectivity can make standard online implementations impractical. Yet other aspects can be easier: raters are often more accessible because organizations tend to be smaller and flatter, meaning the person responsible for hiring or screening is typically on-site and reachable. The prevalence of small firms and decentralized credit markets expands the pool of potential participants. Moreover, because many such firms face large applicant pools but have limited tools to assess candidates, the prospect of receiving curated referrals often makes participation attractive. In practice, recruiting raters—often the most difficult element of IRR studies in high-income settings—may be considerably less challenging in poor countries.

I organize the practical guidance around three steps: (1) *pre-implementation design choices*, (2) *fieldwork execution*, and (3) *post-implementation matching and follow-up*.

Before implementation, researchers must determine how to assemble a pool of candidates, construct realistic application profiles, choose randomization strategies, and select outcome measures. The next subsections address each of these components in turn.

## 5.1 Pre-Implementation Design Choices

### 5.1.1 Assembling the Applicant Pool

The first step in implementing an incentivized ratings design is assembling a pool of real candidates to whom raters would like to be referred. The credibility of the incentive mechanism—and thus the entire design—hinges on the relevance of this pool.

In their study Kessler et al. (2019) drew on an existing résumé bank at the University of Pennsylvania. Such repositories, however, rarely exist in developing-country contexts. More generally, they may not be available for the specific population under study or may not be accessible to researchers, especially early-career ones. This constraint is particularly salient in informal labor markets, where applicants may have limited, heterogeneous, or

poorly documented educational histories, making standardized résumé construction difficult. Researchers have relied on several strategies to overcome these barriers.

**Administrative data.** One approach is to build the applicant pool using administrative data from training institutes, schools, job placement programs, banks, or other organizations that maintain structured records. For example, Macchi and Raisaro (2025) drew on administrative records from vocational training institutes; similar strategies appear in Colonnelli et al. (2022) and Bartoš et al. (2024). When accessible, these records provide standardized information and a natural channel for referrals.

A practical advantage is that this method is cost-effective and ensures that candidate histories are verifiable. In settings where records are not yet digitized, researchers can offer to digitize the database or archival system, creating value for the institution and increasing the likelihood of cooperation.

Tips:

- Confirm that the administrative data align with the attributes decision makers care about; brief surveys can fill gaps.
- Verify that the institution is authorized to share the data. If not, obtain individual-level consent before use.

**Own survey data.** When administrative records are unavailable, researchers can identify the target population, draw a random sample, and collect the information needed to construct application profiles. The key constraint is ensuring that the individuals surveyed are actual prospective candidates (e.g., active job seekers in hiring studies).

Although this approach is more costly, it provides flexibility in what information is collected. If additional supply-side variables are needed—such as beliefs, expectations, or search behavior—they can be incorporated into the same survey. To avoid deception and meet ethical standards, respondents must consent both to data collection and to having their in-

formation shared with raters. I recommend eliciting the latter consent after the survey to avoid strategic behavior.

An alternative, more cost-effective variant is the one implemented by Abebe et al. (2021), who recruited job seekers by advertising positions through newspapers. Interested candidates phoned in for more information, and enumerators collected demographic and background data during these calls, generating a pool of real applicants.

Tips:

- Ensure the survey collects all attributes that raters consider relevant. When possible, request actual application forms used by local firms or financial institutions.
- Consider separating consent for data collection from consent to share information with raters; asking for sharing permission up front may induce strategic reporting.

**Publicly available data.** In some contexts, researchers can construct applicant pools from publicly available sources. Colonnelli et al. (2024b) use procurement tender data published in Ugandan newspapers. Chen (2025) scrape applications from a popular medical crowdfunding platform. Job postings, professional directories, dating apps, and sector-specific registries can serve similar roles depending on the setting.

**Other sources.** Researchers have also developed creative approaches to assembling applicant pools. When administrative records contain relevant information but not in a format suitable for standardized profiles—a common situation in low-income countries—large language models (LLMs) can help translate raw text into usable résumés or application summaries. Bohren et al. (2025) use this approach to generate standardized recommendation letters for fictitious STEM job candidates in the United States.

Another interesting approach is to actively create the applicant pool. Ndayikeza (2025) hired a training firm to teach graduating bachelor students how to create résumés. This training covered the basic rules of making a good résumé, writing cover letters, and succeed-

ing in job interviews. The resulting résumés then formed the basis for the applicant pool, ensuring that profiles reflected real candidates with standardized, professionally formatted applications. Chen (2025) uses the scraped medical crowdfunding platform data to train a large language model to generate realistic hypothetical calls for donations.

### 5.1.2 Randomization of Traits and Application Content

Once candidates are identified, researchers must translate their information into standardized application profiles. This typically requires harmonizing the raw data, deciding which attributes to display, and ensuring that the resulting profiles resemble the materials raters normally encounter. In settings where applicants rarely submit formal résumés, profile templates may need to be simplified, reformatted, or adapted to local conventions.

The original incentivized ratings design relied on résumé "shells" where every field was randomized dynamically in Qualtrics. This means that hypothetical résumés were not created ex ante. Rather, the authors created several options for each field (e.g., GPA, major, work experience), and combinations of these options were randomly selected to create a unique résumé for each evaluation. This is an ambitious approach that allows for high variability in profiles. However, if researchers rely on an offline platform like SurveyCTO for data collection, this approach may not be feasible.

If the focal attribute is binary or categorical (e.g., male/female, high/low credit scores), pre-generating randomized application shells can simplify fieldwork. The approach works as follows: first, craft a set of realistic candidate profiles by cross-randomizing non-treatment attributes using statistical software. Then, for each profile, create different versions corresponding to the treatment conditions to be tested.

Beyond the feasibility, another benefit of this approach, relative to the standard KLS (2019) implementation, is that it allows researchers to study differential treatment based on group identity while holding fixed all other characteristics. Indeed, this within-application randomization yields more precise estimates with application-level fixed effects.

A key design consideration is how closely the distribution of focal attributes in the hypothetical applicant pool should mirror the real distribution in the population. Researchers face a trade-off: increasing realism can reduce statistical power, for example, in a situation where one is interested in testing for discrimination of a minority identity group; at the same time, keeping a distribution closer to the real distribution in the population may also mitigate experimenter-demand concerns.

### 5.1.3 Using Photos and Other Rich Media

In many settings, researchers may wish to include photos or other rich media in application profiles. This is particularly relevant in markets where decision-makers normally see candidates' faces before screening applicants, such as in frontline hiring, informal labor markets, small-firm recruitment, or in-person credit evaluations. Images may also be necessary when the attributes of interest concern physical characteristics (e.g., gender presentation, body size, attire, or religiosity).

Using photos introduces several challenges. First, photos bundle multiple traits—attractiveness, expression, age cues, and more—making it difficult to isolate the attribute being experimentally varied. Second, images raise ethical considerations: consent is needed both from real individuals whose photos are used and from candidates represented by those photos. Third, the information conveyed by the photo must be consistent with the underlying profile; mismatches between appearance and résumé content can reduce credibility and increase noise.

Researchers have developed several strategies to address these issues. When possible, keeping appearance constant across treatment conditions is the cleanest/simplest way to avoid confounds. This strategy works well when the treatment occurs in text fields (e.g., qualifications, work history) and the photo does not need to vary across conditions. Obviously, this is infeasible when the treatment itself requires different photos (e.g., gender experiments).

When different photos across treatments are necessary (e.g., because of attributes like gender or race), researchers must ensure that the photos do not introduce confounding information. The standard approach involves having raters evaluate photos on relevant characteristics (e.g., attractiveness, professionalism, trustworthiness) to select photos that are balanced on these outcomes. For example, Low (2024) purchased stock photos similar to those on dating websites and had undergraduates rate each photo's attractiveness and guess the individual's age. Photos can then be grouped into sets that are statistically balanced across experimental treatments. An alternative approach is to blur photographs to preserve realism while removing specific confounds. Sun and Huo (2020) blur photos to mask gender while Chan (2022) blur photos to convey doctors' race while eliminating attractiveness and gender confounds.

A different set of challenges arises when photographs are used to deliver the treatment itself. In these cases, researchers may rely on digital manipulation along the treatment dimension. For example, to randomize body size, Macchi (2023) photographed Kampala residents with consent and hired a professional photographer to digitally alter weight-related appearance. Independent raters screened manipulated images for realism, flagging and excluding those that appeared unnatural. To maintain consistency, only the manipulated versions were used (not the originals), ensuring that all profiles reflected a comparable level of digital editing. A separate validation exercise confirmed that the final images corresponded to the intended perceived BMI categories.

Another approach is that of Fiorin, La Ferrara and Shofia (2022) studying the labor market implications of religious markers in Indonesia. A professional photographer recruited non-professional models and took baseline photos without religious markers (i.e., no headscarf or beard). A professional editor then digitally added either a headscarf or a beard to create a matched variant, yielding a pair of images per model, one without and one with a religious marker. The style and color of headscarves and the length and style of beards varied across individuals to reflect natural variation.

Finally, images may also need to represent multidimensional objects rather than people, such as neighborhoods, products, or investment opportunities. For example, Ferlenga (2025) use collages of city photographs to study preferences over urban amenities: each collage shows standard features of a mid-sized U.S. city, and a single image is randomized to display either a Confederate monument or a neutral scene.

Tips:

- Pilot the full set of profiles with a small sample of local raters to identify strange photos, if photo-manipulated.

### 5.1.4 Choosing Outcomes and Evaluation Formats

A key design choice in incentivized ratings studies concerns the type of outcome researchers elicit from decision makers. The optimal format depends on the setting, the literacy of raters, and the conceptual question of interest.

**Binary outcomes.** Binary outcomes—such as "Would you interview this candidate?" or "Would you approve this loan?"—are straightforward, easy to explain, and particularly useful in low-literacy environments or time-constrained field settings. They map directly onto real decisions and reduce cognitive load. Their main limitation is granularity: binary choices provide coarse variation and may mask meaningful differences in preferences across profiles.

**Likert or numerical ratings.** Numerical rating scales (e.g., 1–5 or 1–10) allow for richer variation and typically generate higher statistical power. They are well suited to contexts where raters are accustomed to evaluating candidates along multiple dimensions. However, ratings are more cognitively demanding and may lead to heaping or invariant use of the scale. One note is that making decision-makers think about different aspects of the application, for example, by having them rate on various Likert outcomes, can increase attribute salience and, in turn, may induce experimenter demands.

**Rankings and conjoint formats.** Some IRR applications ask raters to rank multiple profiles or choose among bundled attributes, in practice following a conjoint design. These formats are powerful for studying trade-offs across attributes, and in some contexts, like the doctors' discrimination in Zhang (2024), they offer a more natural setup. The main drawback is complexity: rankings and conjoint tasks are more difficult to explain and may be unsuitable in high-pressure or low-literacy environments. In practice, also, these ranking outcomes are often converted to binary comparisons for analysis, suggesting that the added complexity may not yield substantive benefits.

**When continuous measures are natural.** A different context arises when the outcome itself is a meaningful continuous variable. For example, in studies where respondents allocate resources across candidates (Colonnelli et al., 2024a) or assign wages, the continuous measure is naturally preferred because it captures economically meaningful variation in the intensity of preferences or willingness to invest.

Tips:

- If using a binary outcome, clearly explain that the total number of matches is fixed so that respondents do not mistakenly expect that approving more candidates increases their chances of receiving a referral.
- To gain insight into decision-making processes, researchers can elicit both binary decisions and secondary ratings (e.g., reliability, creditworthiness). However, collecting many secondary outcomes increases respondent burden, creating a trade-off with the number of profiles evaluated.

## 5.2 Implementation: Fieldwork

Once the design choices are finalized, the next stage is executing the incentivized ratings study in the field. This requires recruiting raters, implementing the randomization proto-

col, and managing evaluation sessions. The subsections that follow discuss each of these components in turn.

### 5.2.1 Recruiting Decision Makers

First, researchers must identify and engage the individuals who actually make screening or hiring decisions. In many developing-country contexts, organizations are smaller and flatter than in high-income settings, which often makes it easier to reach the relevant decision maker directly. At the same time, ensuring that respondents have real decision authority is crucial for the credibility of both the task and the incentive mechanism.

In practice, recruitment often resembles a standard in-person survey: researchers visit workplaces or institutional branches, introduce the study as an opportunity to access a curated pool of candidates, and invite decision makers to participate. Two implementation strategies have been especially common.

When a complete population directory exists—typically assembled from administrative data—researchers can sample directly from it. For example, Macchi (2023) obtained a registry of loan officers from the Ugandan Ministry of Finance, complete with phone numbers and office locations. The advantage of such a list is that it allows researchers to calculate response rates and assess the representativeness of participating decision makers.

When no directory or list is available, researchers can conduct a listing exercise to identify potential raters, as in Macchi and Raisaro (2025). This approach involves brief visits or phone calls to firms, branches, or community organizations to determine who is responsible for hiring, screening, or lending decisions. Although more time-consuming, listing exercises can generate better sampling in informal markets.

IRR also makes it possible to conduct well-powered experiments with small but strategically important populations. Because each decision maker evaluates many profiles, researchers can achieve high statistical power even when the relevant population is small. For example, Colonnelli et al. (2024a) successfully implemented an incentivized ratings design

with entrepreneurs and venture capitalists in sub-Saharan Africa. Similar approaches could be applied in other elite decision-making environments, including political appointments, executive hiring, or specialized credit markets.

Tips:

- Verify that the decision maker has actual authority to select candidates for the positions of interest and has an actual need for the candidates.

### 5.2.2 Evaluations: Number and Approach

Deciding how many evaluations each rater should complete requires balancing two competing considerations: statistical power and internal data quality. More evaluations per rater increase precision and reduce the total sample size needed to detect treatment effects. However, as the number of profiles grows, respondents may become fatigued, potentially compromising data quality. This trade-off is particularly salient in studies of discrimination. As noted in Litwin and Low (2025), evidence from Kessler et al. (2019), for example, shows that implicit biases are more likely to surface when respondents are fatigued. This raises a conceptual question about whether discrimination measured under fatigue reflects the behavior of interest or an artifact of the experimental design. Keeping the number and format of evaluations close to what raters encounter in their daily work helps mitigate both internal and external validity concerns.

The ideal number of evaluators per rater, therefore, is context specific. Across studies reviewed in this paper, the number of profiles evaluated ranges from 2 to 50, with a median of about 20 and a mean of 18.8. An advantage of in-person enumeration is that enumerators can help sustain respondents' attention for longer periods, allowing decision makers to rate more applications. For example, unlike the online implementation in Kessler et al. (2019), both Macchi (2023) and Macchi and Raisaro (2025) conducted in-person sessions in which decision makers evaluated more than 20 CVs with no evidence of order effects.

Relevant factors to consider include on the complexity of the materials, the familiarity with the evaluation format, and the strength of incentives. Tablets or online interfaces may feel unfamiliar to decision makers who typically review paper CVs or make in-person assessments. Highly detailed or information-rich profiles may induce fatigue, while overly simplified profiles may appear unrealistic or fail to provide the information needed for coherent decisions. Professional decision-makers with stronger economic incentives can typically rate more candidates than laypeople—a consideration that affects both statistical power (more ratings per respondent) and data quality (sustained attention throughout the task).

Tips:

- Randomize profile order and check for order effects in pilots.
- Match profile design to real-world practices: see Appendix Figure 4 (Macchi, 2023).

### 5.2.3 Delivering Instructions and Ensuring Comprehension

A key challenge in implementing incentivized ratings designs is ensuring decision makers understand how their choices translate into outcomes. The clearer the incentives, the more precise our measurements of preferences and treatment effects will be. Emphasizing that the time and effort they invest in the selection process will improve the quality of the match is important.

At the same time, explanations must avoid revealing the study's hypotheses or unintentionally shifting respondents' attention toward specific attributes. The goal is to provide enough information for comprehension while preserving the natural decision environment. For example, Macchi (2023) and Zhang and Zou (2025) emphasize the general purpose of their matching tool while avoiding mention of specific research objectives (e.g., weight discrimination or ESG preferences) to minimize Hawthorne effects.

Enumerators can introduce the exercise by framing it as a matching or referral service. This framing is especially effective in labor and credit markets where referrals are already common and where employers or loan officers regularly evaluate informal candidate pools.

Presenting the task as an opportunity to receive a curated set of high-quality candidates helps raters understand the practical value of the exercise and encourages careful, engaged evaluations. Field experience suggests that emphasizing the direct benefits of participation increases raters' attention.

Another delicate issue is communicating the hypothetical nature of the profiles. Researchers must avoid deception, yet overemphasizing that profiles are hypothetical can make the task feel artificial or game-like. A useful strategy is to emphasize that while profiles are constructed to ensure comparability, the referrals they generate will involve real individuals matched to the respondent's preferences. In-person sessions allow enumerators to detect confusion in real time and reinforce the credibility of the referral mechanism. Professional formatting of application materials—printed CVs, standardized layouts, or realistic cover sheets—can also help keep the task grounded.

Different populations require different instructional tools. In high-literacy environments, a clear verbal explanation may suffice, as in Nielsen and Rigotti (2022), who provided a detailed walkthrough of how preferences would be inferred and used. In lower-literacy contexts, more structured tools improve comprehension. Colonnelli et al. (2024b), for example, used an animated video shown during in-person visits to ensure consistent explanation quality across enumerators.[3]

Finally, when concern about hypothetical framing is especially acute, one alternative is to incorporate real applicants. However, doing so reduces experimental control and limits the ability to randomize profile attributes. Abebe et al. (2021) used real CVs and skill tests, but this approach sacrifices the full flexibility of the IRR design. A hybrid strategy, as in Raisaro (2023), includes one real candidate among hypothetical profiles and implements the matching mechanism only for the real candidate, maintaining credibility while preserving some experimental variation.

Tips:

[3]Few of the papers we reviewed clarify how they explained the mechanism to participants, making it difficult to assess best practices or learn from others' experiences.

- Use simple visual or verbal demonstrations to explain the matching algorithm.
- Use comprehension checks before starting the task.
- Train enumerators to monitor and address confusion and fatigue in real time.

## 5.3 Post-Implementation: Implementing Referrals

The original design activates matches by providing raters with information about selected applicants. However, how the match is activated should align with contextual practices and may need to be reversed depending on context. This is because the likelihood of a match happening can affect the incentives of the decision makers.

For example, in markets where candidates typically present themselves to be selected (e.g., loan applicants in Macchi, 2023, spot labor market Breza and Kaur, 2025), this approach will not work; indeed, matches will be unlikely to materialize because candidates are expected to initiate contact. In such contexts, an alternative is providing selected applicants with the bank or loan officer's details, precise directions, and a letter of introduction. Importantly, we had obtained consent from loan officers to share their contact information with applicants.

How to deliver referral information involves trade-offs between cost, reliability, and credibility. Mailing or e-mailing is often impractical in low-income contexts for the same reasons it faces limitations in correspondence studies. Text messages are low-cost but may be taken as scams depending on content. In those cases, phone calls provide the most credibility and allow for immediate clarification of questions, but information may be forgotten or misunderstood. A combined approach— an initial phone call followed by a text message with key details—can be a good way to balance these trade-offs.

# 6 Ethical and Conceptual Considerations

## 6.1 Non-Deception and Commitments to Participants

A defining feature of incentivized ratings designs is the commitment to avoid deception. This requires that any promises made to participants—especially regarding the referral mechanis—are feasible and credible. For example, when the incentive involves access to a curated pool of candidates, it may be best not to guarantee a specific number of referrals if, e.g., candidate no-shows are common in that market.

## 6.2 Matching Algorithms and Discriminatory Preferences

Building on the principle of non-deception discussed above, when incentivized ratings designs are used to study discrimination, researchers face a tension between maintaining incentive compatibility and avoiding the amplification of discriminatory preferences. If raters express biased preferences—for example, a preference not to hire women—should the matching algorithm honor these preferences to preserve the credibility of the incentive, or should it override them to avoid perpetuating discrimination?

Kessler et al. (2019) argues for the latter approach: when preferences are expressed over legally or ethically protected characteristics, the algorithm should not match on these dimensions. Because the goal of the referral mechanism is to provide high-quality candidates rather than to replicate every element of stated preferences, the algorithm can prioritize other substantive attributes such as skills, experience, or reliability. Under this approach, respondents continue to receive candidates who are well-matched on meaningful, job-relevant characteristics, while the research design avoids complicity in discriminatory behavior. The underlying principle is that as long as recommended candidates reflect the non-discriminatory determinants of job performance, referrals remain valuable and the design remains non-deceptive, even if they do not align perfectly with biased preferences.

Only 7 of the 13 discrimination studies that we identify in this review explicitly discuss the matching algorithm, while 6 do not discuss their choice. Three of these seven studies choose not to match on the discriminatory variable. The alternative approach is to implement the matching based on all the variables, including potentially discriminatory ones.

## 6.3 Experimenter Demands

A concern in the experiment is that respondents may try to infer what the researcher wants them to choose and adjust their responses accordingly. This is particularly worrisome in experiments that aim at measuring sensitive preferences, like discrimination. In the incentivized ratings design, this concern is mitigated by the fact that respondents have real stakes in receiving good matches. However, it remains important to design the experiment in a way that minimizes such demands. For example, as discussed above, in settings where there is a clearly socially desirable set of preferences, one may want to avoid explicitly stating the goal of the study.

One can also implement simple strategies to assess whether respondents perceive experimenter demand and whether it influences their choices. For example, inspired by De Quidt et al. (2018), researchers can ask respondents directly about what they believe the research team wants them to choose. However, the framing and timing of this question matter. It is advisable to ask this question at the category level rather than for each individual profile and after the elicitation procedure. For example, in Macchi and Raisaro (2025), we ask employers, "Do you think we (the research organization) would be happier if you hired (1) men, (2) women, or (3) we are indifferent?"

Erda and Shrader (2023), which studies how race biases appearance-related beliefs in academia, uses a complementary approach—also inspired by De Quidt et al. (2018)—to test for experimenter demand effects. They divide their treatment group into three subgroups, each receiving a different degree of nudging toward the hypothesized direction of demand, ranging from a mild prompt ("Previous respondents tended to choose X.") to a stronger one

("This study examines whether respondents choose X."). The difference in responses across these subgroups and the no-demand subgroup provides an effective bound on the magnitude of experimenter demand.

## 6.4 External Validity and Selection of Raters

An important consideration for incentivized ratings designs is the potential for selection into participation. Because IRR studies provide access to a curated pool of candidates, raters who opt into the study may differ systematically from the broader population of decision makers. In particular, individuals or organizations facing more severe hiring frictions, credit officers with heavier caseloads, or employers who struggle to recruit through standard referral channels may value the matching opportunity more and therefore be more likely to participate.

These selection patterns can affect external validity. If participants disproportionately represent employers with high labor demand, weak referral networks, or greater openness to trying new recruitment mechanisms, estimated treatment effects may not generalize to all firms or decision makers. Documenting observable differences between participants and non-participants—such as firm size, sector, hiring volume, or staffing constraints—can help clarify the extent of this concern.

The severity and implications of selection vary by context. In many low- and middle-income settings, where referral-based hiring is the norm and formal recruitment channels are limited, selection into IRR studies may be particularly pronounced. At the same time, conducting recruitment in person allows researchers to reach a broader and potentially more representative sample of decision makers than would be possible through online or opt-in recruitment alone. When feasible, researchers should leverage this flexibility to construct samples that reflect the diversity of the underlying hiring or lending environment.

# 7 Data Considerations

## 7.1 Pre-registration and data cleaning

A preregistration must clearly define the data-generating process, beginning with the unit of randomization and the unit of observation (the profile–rater evaluation). Several studies in our review did not report basic design choices such as the number of profiles evaluated per rater.

A preregistration should also define the primary and secondary outcomes, especially because IRR studies routinely elicit more than one evaluation per profile (e.g., a binary accept/reject decision, Likert ratings, trait-specific assessments, or rankings). When outcomes will be transformed—for example, converting rankings or continuous scores into binary indicators—this should be declared in advance.

Because incentivized ratings studies are sensitive to low-quality responses, preregistration should specify exclusion rules for inattentive or non-discriminating raters, such as individuals who accept all profiles, reject all profiles, or give identical scores to every application. In settings where respondents may skip through applications, researchers may want to prespecify a minimum number of completed evaluations for inclusion.

Preregistrations can also document how matches will be implemented: whether referrals are delivered to raters or candidates, how communication will occur (e.g., phone plus SMS), and the criteria for considering a referral "delivered."

Finally, when multiple attributes are randomized within profiles, preregistration should clearly state which hypotheses will be tested (e.g., focus on gender or race discrimination) and which randomized attributes serve as controls, given that often multiple attributes are cross-randomized within profiles.

### 7.2 Data Analysis

According to Abadie et al. (2023), one should consider clustering standard errors when treatment assignments or sampling are correlated. Since each rater contributes multiple evaluations to the dataset, standard practice is to cluster standard errors at the rater level if decision-makers were sampled from some super-population of interest. Clustering at the decision-maker level is particularly important when decision-makers are assigned to different treatments.

In conjoint designs, where treatments are randomized at the level of the choice set rather than the individual profile, standard errors are typically clustered at the choice set level to account for within-task correlation across profile ratings.

Among the papers we review, standard error clustering practices vary: 22 of 37 studies (58%) cluster at the rater level, while 16 studies use alternative clustering approaches or do not cluster standard errors at all. Among the 13 published papers, only 5 papers do not cluster at the rater level despite each rater providing multiple (mean 18.6) observations.

Including rater fixed effects in the main specification absorbs all between-rater heterogeneity and increases precision by controlling for individual-level rating tendencies. When the design presents multiple versions of the same profile (e.g., manipulating gender or race only), profile fixed effects can further increase statistical power through within-shell comparisons.

## 8 Conclusion

The incentivized ratings design provides a powerful and practical way to uncover preferences—and, when relevant, detect discrimination—in settings where decisions are made in person. By framing the exercise as a referral system in which respondents evaluate profiles and receive matches based on their choices, the approach maps naturally onto environments where informal networks and referrals dominate, particularly in low-income countries. Be-

cause each rater evaluates many applications, IRR designs generate high statistical power at relatively low cost, even when studying small but strategically important populations.

The design is also highly flexible. Researchers can embed randomized variations in profile attributes, introduce information treatments to test mechanisms, or elicit beliefs to study the foundations of discriminatory behavior. These features make IRR a valuable addition to the development economist's experimental toolkit. When carefully implemented, it offers a scalable and credible method for studying decision makers preferences and discrimination in poor-country settings.

# References


**Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge**, "When should you adjust standard errors for clustering?," *The Quarterly Journal of Economics*, 2023, *138* (1), 1–35.

**Abebe, Girum, Stefano A. Caria, and Esteban Ortiz-Ospina**, "The Selection of Talent: Experimental and Structural Evidence from Ethiopia," *American Economic Review*, 2021, *111* (6).

**Alabrese, Eleonora, Francesco Capozza, and Prashant Garg**, "Politicized Scientists: Credibility Cost of Political Expression on Twitter," Technical Report, CAGE Research Centre Working Papers 2024.

**Archibong, Belinda, Francis Annan, Anja Benshaul-Tolonen, Oyebola Okunogbe, and Ifeatu Oliobi**, "Firm Culture: How Social Norms Affect Gender Bias in Hiring in Online Labor Markets," Technical Report, Centre for Economic Policy Research Working Papers 2025.

**Banerjee, Abhijit, Marianne Bertrand, Saugato Datta, and Sendhil Mullainathan**, "Labor market discrimination in Delhi: Evidence from a field experiment," *Journal of comparative Economics*, 2009, *37* (1), 14–27.

**Bartŏs, Vojtěch, Silvia Castro, Kristina Czura, and Timm Opitz**, "Gendered Access to Finance: The Roles of Team Formation, Idea Quality, and Implementation Constraints in Business Evaluations," Technical Report, CESifo Working Paper 2024.

**Beam, Emily A, Asad Islam, Joshua D Merfeld, and Naveen Wickremeratne**, "Improving Gender Norms in the Workplace," 2024.

**Beaman, Lori and Jeremy Magruder**, "Who gets the job referral? Evidence from a social networks experiment," *American economic review*, 2012, *102* (7), 3574–3593.

**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*, 2004, *94* (4), 991–1013.

**Bohren, J Aislinn, Peter Hull, and Alex Imas**, "Systemic discrimination: Theory and measurement," *The Quarterly Journal of Economics*, 2025, *140* (3), 1743–1799.

**Breza, Emily and Supreet Kaur**, "Labor Markets in Developing Countries," *Annual Review of Economics*, 2025, *17*.

**Bursztyn, Leonardo, Bruno Ferman, Stefano Fiorin, Martin Kanz, and Gautam Rao**, "Status Goods: Experimental Evidence from Platinum Credit Cards*," *The Quarterly Journal of Economics*, 12 2017, *133* (3), 1561–1595.

**Bustelo, Monserrat, Ana Maria Diaz, Jeanne Lafortune, Claudia Piras, Luz Magdalena Salas, and José Tessada**, "What is the price of freedom? Estimating women's willingness to pay for job schedule flexibility," *Economic Development and Cultural Change*, 2023, *71* (4), 1179–1211.

**Carpio, Lucia Del and Thomas Fujiwara**, "Do gender-neutral job ads promote diversity? Experimental evidence from Latin America's tech sector," Technical Report, National Bureau of Economic Research 2023.

**Cefala, Luisa, Supreet Kaur, Heather Schofield, and Yogita Shamdasani**, "Habit Formation in Labor Supply," Technical Report, Working paper 2024.

**Chan, Alex**, "Discrimination against doctors: A field experiment," *Unpublished manuscript*, 2022.

**Chen, Junhao**, "Is Deservingness Merit-based or Need-based? College Selectivity and Medical Crowdfunding Outcomes," *College Selectivity and Medical Crowdfunding Outcomes (April 20, 2025)*, 2025.

**Cole, Shawn, Martin Kanz, and Leora Klapper**, "Incentivizing calculated risk-taking: Evidence from an experiment with commercial bank loan officers," *The Journal of Finance*, 2015, *70* (2), 537–575.

**Colonnelli, Emanuele, Bo Li, and Ernest Liu**, "Investing with the government: A field experiment in China," *Journal of Political Economy*, 2024, *132* (1), 248–294.

**_ , Francesco Loiacono, Edwin Muhumuza, and Edoardo Teso**, "Do information frictions and corruption perceptions kill competition? a field experiment on public procurement in uganda," 2024.

**_ , Marcio Cruz, Mariana Pereira-Lopez, Tommaso Porzio, and Chun Zhao**, "Startups in Africa," 2025. Working paper, coming soon.

**_ , Tim McQuade, Gabriel Ramos, Thomas Rauter, and Olivia Xiong**, "ESG Is the Most Polarizing Nonwage Amenity: Evidence from a Field Experiment in Brazil," in "AEA Papers and Proceedings," Vol. 115 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2025, pp. 146–152.

_ , **Valdemar Pinho Neto, and Edoardo Teso**, "Politics at work," Technical Report, National Bureau of Economic Research 2022.

**Duflo, Esther and Marianne Bertrand**, "Field Experiments on Discrimination," in Abhijit V. Banerjee and Esther Duflo, eds., *Handbook of Field Experiments*, Vol. 1, Elsevier, 2016, pp. 309–393.

**Ebrahimian, Mehran and Ye Zhang**, "Matching and Bargaining in Entrepreneurial Finance: Evidence From an Experimental System," Technical Report, SSRN Working Papers 2025.

**Erda, Tarikua and Jeffrey Shrader**, "Appearance Norms in Academic Settings," AEA RCT Registry 2023. Registered October 17, 2023.

**Exley, Christine L, Raymond Fisman, Judd B Kessler, Louis-Pierre Lepage, Xiaomeng Li, Corinne Low, Xiaoyue Shan, Mattie Toma, and Basit Zafar**, "Information-optional policies and the gender concealment gap," Technical Report, National Bureau of Economic Research 2024.

**Feng, Junlong, Ofir Gefen, Ye Zhang, and Weijie Zhong**, "Statistical Discrimination in Two-sided Matching Markets: Experimental and Theoretical Evidence," 2024.

**Ferlenga, Francesco**, "Symbols of Oppression: The Role of Confederate Monuments in the Great Migration," Technical Report, SSRN Working Papers 2025.

**Fiorin, Stefano, Eliana La Ferrara, and Naila Shofia**, "Labor Markets Matching in Indonesia: Estimating Employers' Preferences with Incentivized Resume Rating," AEA RCT Registry June 2022. June 23.

**Fisman, Raymond, Daniel Paravisini, and Vikrant Vig**, "Social proximity and loan outcomes: Evidence from an Indian Bank," *American Economic Review*, 2011.

**Galarza, Francisco B and Gustavo Yamada**, "Labor market discrimination in Lima, Peru: Evidence from a field experiment," *World Development*, 2014, *58*, 83–94.

**Gallen, Yana and Melanie Wasserman**, "Does information affect homophily?," *Journal of Public Economics*, 2023, *222*, 104876.

**Garz, Seth, Xavier Giné, Dean Karlan, Rafe Mazer, Caitlin Sanford, and Jonathan Zinman**, "Consumer protection for financial inclusion in low-and middle-income countries: Bridging regulator and academic perspectives," *Annual Review of Financial Economics*, 2021, *13* (1), 219–246.

**Gentile, Elisabetta, Nikita Kohli, Nivedhitha Subramanian, Zunia Saif Tirmazee, and Kate Vyborny**, "Barriers to entry: Decomposing the gender gap in job search in urban Pakistan," Technical Report, ADB Economics Working Paper Series 2023.

**Heath, Rachel**, "Why do firms hire using referrals? Evidence from Bangladeshi garment factories," *Journal of Political Economy*, 2018, *126* (4), 1691–1746.

**_ , Arielle Bernhardt, Girija Borker, Anne Fitzpatrick, Anthony Keats, Madeline McKelway, Andreas Menzel, Teresa Molina, and Garima Sharma**, "Female labour force participation," *VoxDevLit*, 2024, *11* (1), 1–43.

**Heckman, James J**, "Detecting discrimination," *Journal of economic perspectives*, 1998, *12* (2), 101–116.

**Kessler, Judd B, Corinne Low, and Colin D Sullivan**, "Incentivized resume rating: Eliciting employer preferences without deception," *American Economic Review*, 2019, *109* (11), 3713–3744.

**Litwin, Ashley and Corinne Low**, "Measuring Discrimination with Experiments," 2025.

**Low, Corinne**, "Pricing the biological clock: The marriage market costs of aging to women," *Journal of Labor Economics*, 2024, *42* (2).

_ **et al.**, "Estimating Employer Preferences Over Workers in India," AEA RCT Registry October 2023. October 17.

**Macchi, Elisa**, "Worth your weight: experimental evidence on the benefits of obesity in low-income countries," *American Economic Review*, 2023, *113* (9), 2287–2322.

_ **and Claude Raisaro**, "Hidden Gender Discrimination," 2025.

**Martin, Diego**, "Women Seeking Jobs with Limited Information: Evidence from Iraq," Technical Report, Harvard Growth Lab Working Papers 2024.

**Maurer-Fazio, Margaret**, "Ethnic discrimination in China's internet job board labor market," *IZA Journal of Migration*, 2012, *1* (1), 12.

**Ndayikeza, Michel Armel**, "Underemployment of college graduates: is doing anything better than doing nothing?.," *Journal of Development Economics*, 2025, *174*.

**Nielsen, Kirby and Luca Rigotti**, "Revealed incomplete preferences," *arXiv preprint arXiv:2205.08584*, 2022.

**Park, Hyoeun**, "Patterns of Discrimination and the Effect on Model Minorities," 2025.

**Quidt, Jonathan De, Johannes Haushofer, and Christopher Roth**, "Measuring and bounding experimenter demand," *American Economic Review*, 2018, *108* (11), 3266–3302.

**Raisaro, Claude**, "Incentives Justifying Nonconformity: Experimental Evidence from Motortaxi Organizations in Uganda," 2023.

**Rosenzweig, Mark R**, "Labor markets in low-income countries," *Handbook of development economics*, 1988, *1*, 713–762.

**Sinha, Sourav and Zhengren Zhu**, "Labor Market Returns to Upskilling - A Combination of Audit Study and Resume Review," AEA RCT Registry April 2020. April 21.

**Spadavecchia, Lorenzo, Paola Antonia Profeta, and Maddalena Ronchi**, "Managerial implicit stereotypes and where to find them: Evidence from Incentivized Resume Rating," AEA RCT Registry February 2023. February 01.

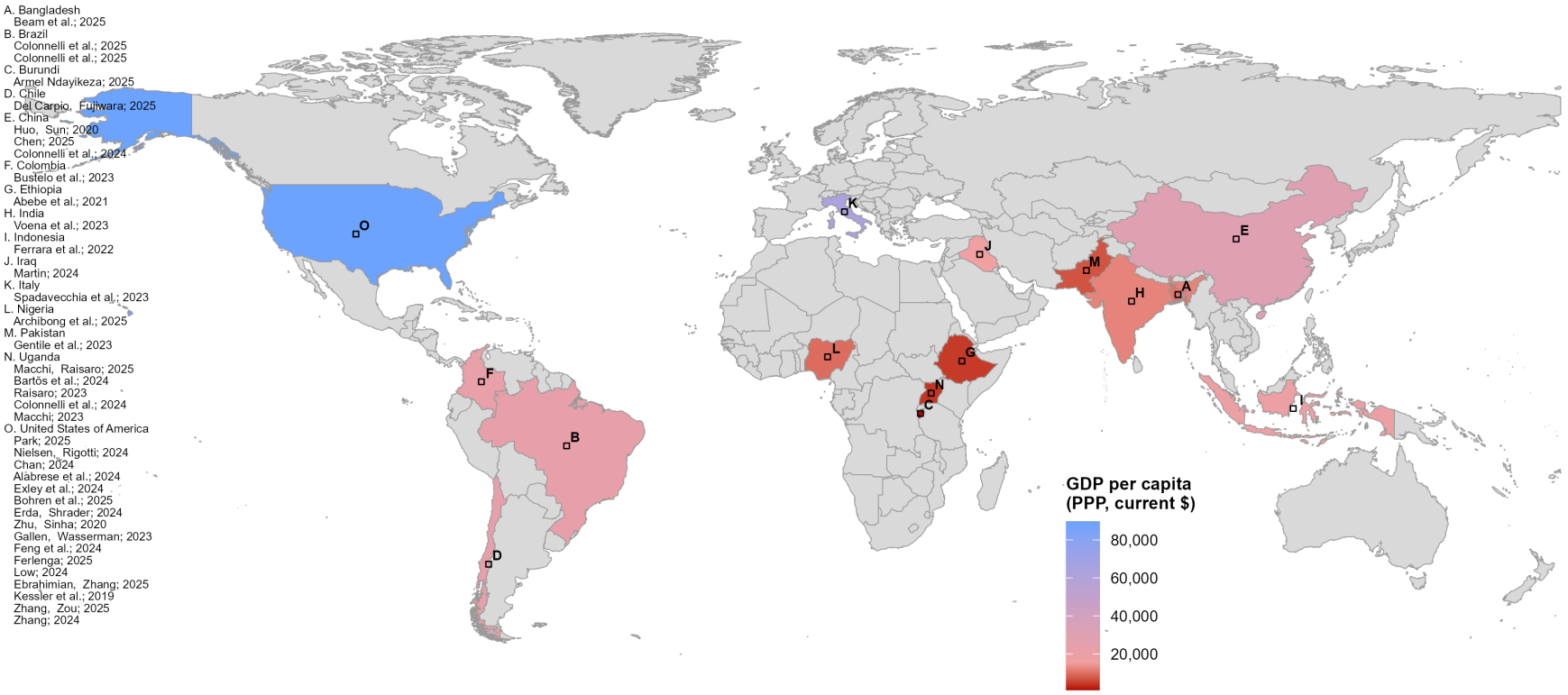
**Sun, Lihua and Yuezhou Huo**, "The Return to Foreign and Domestic Master's Degrees in the Chinese Labor Market," in "2020 APPAM Fall Research Conference" APPAM 2020.

**Zhang, Ye**, "Discrimination in the Venture Capital Industry: Evidence from Two Randomized Control Trials," *Management Science*, 2024.

_ **and Eric Zou**, "ESG Aversion: Experimental Evidence on Perceptions and Preferences," Technical Report, National Bureau of Economic Research 2025.

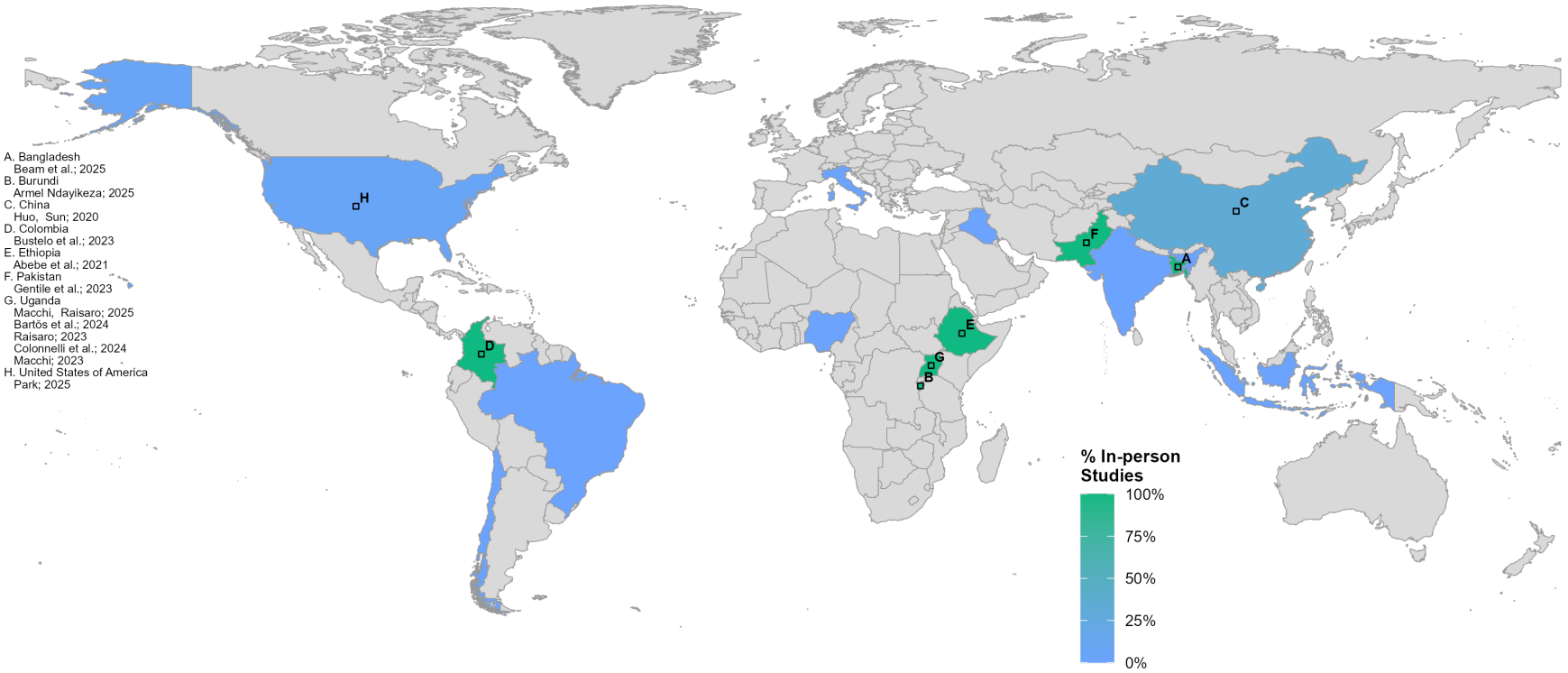
# Figures

Figure 2: Global Distribution of IRR Studies by Country and Income Level




*Note:* This figure maps all incentivized ratings design (IRR) studies identified in our review, by the country in which the study was conducted. Each marker is placed at the country centroid and shaded according to GDP per capita (IMF data).

Figure 3: Share of IRR Studies Conducted In Person Across Countries




*Note:* This figure maps papers that use the incentivized ratings design mechanism by the percent of papers done in person in each country. Each marker is placed around the centroid of the respective country.

# Tables

Table 1: Summary Statistics

| | > 4500 USD per capita | | ≤ 4500 USD per capita | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Twist: Design | 0.44 | 0.5 | 0.75 | 0.5 |
| Implementation | 0.48 | 0.5 | 0.75 | 0.5 |
| Outcome: Likert | 0.48 | 0.5 | 0.25 | 0.5 |
| Continuous | 0.20 | 0.4 | 0.08 | 0.3 |
| Binary | 0.24 | 0.4 | 0.42 | 0.5 |
| In Person vs Online | 0.12 | 0.3 | 0.75 | 0.5 |
| Num. Applications | 20.91 | 14.6 | 17.00 | 13.3 |
| Discrimination | 0.32 | 0.5 | 0.42 | 0.5 |
| Setting: Labor Market | 0.56 | 0.5 | 0.75 | 0.5 |
| Topic: Gender | 0.44 | 0.5 | 0.50 | 0.5 |
| Country Income: High | 0.68 | 0.5 | 0.00 | 0.0 |
| Upper Middle | 0.32 | 0.5 | 0.00 | 0.0 |
| Lower Middle | 0.00 | 0.0 | 0.42 | 0.5 |
| Low | 0.00 | 0.0 | 0.58 | 0.5 |
| Region: Africa | 0.00 | 0.0 | 0.67 | 0.5 |
| Asia | 0.16 | 0.4 | 0.33 | 0.5 |
| Europe | 0.04 | 0.2 | 0.00 | 0.0 |
| North America | 0.64 | 0.5 | 0.00 | 0.0 |
| South America | 0.16 | 0.4 | 0.00 | 0.0 |
| Status: Published | 0.36 | 0.5 | 0.25 | 0.5 |
| Working Paper | 0.52 | 0.5 | 0.58 | 0.5 |
| Registered | 0.12 | 0.3 | 0.17 | 0.4 |
| Year: 2019 | 0.04 | 0.2 | 0.00 | 0.0 |
| 2020 | 0.08 | 0.3 | 0.00 | 0.0 |
| 2021 | 0.00 | 0.0 | 0.08 | 0.3 |
| 2022 | 0.00 | 0.0 | 0.08 | 0.3 |
| 2023 | 0.12 | 0.3 | 0.33 | 0.5 |
| 2024 | 0.40 | 0.5 | 0.17 | 0.4 |
| 2025 | 0.36 | 0.5 | 0.33 | 0.5 |
| Observations | 25 | | 12 | |

*Note:* This table reports summary statistics on key characteristics of each paper, including AEA-registered experiments without an associated working paper, split by high- and low-income countries. All variables are binary except *Number of Applications*. Income classifications follow the World Bank. Papers are classified as design or implementation twists based on deviations from KLS (2019). Data were collected by the authors. A paper is classified as a design twist if it (i) departs from KLS (2019) in the design of the incentive mechanism, (ii) uses IRR to study a channel of discrimination (e.g., by varying information provided to decision-makers), or (iii) embeds IRR within a broader experimental design. A paper is classified as an implementation twist if it uses a conjoint menu of applications, is conducted in person, or departs from KLS (2019) in the randomization procedure.

# Table 2: IRR Literature Review

| Citation<br>Applications # Decision makers | Title | Publication | Region | Country | Income Class. | In person/Online | Twist | JEL Code | Setting | Topic |
|---|---|---|---|---|---|---|---|---|---|---|
| Abebe et al. (2021) | The Selection of Talent: Experimental and Structural Evidence from Ethiopia | American Economic Review | Africa | Ethiopia | Low | In person | Both | J2 | Labor market | Information frictions |
| Alabrese et al. (2024) | Politicized Scientists: Credibility Cost of Political Expression on Twitter | Unpublished | North America | U.S.A | High | Online | Implementation | I2 | Political economy | Political discrimination |
| Archibong et al. (2025) | Firm Culture: How Social Norms Affect Gender Bias in Hiring in Online Labor Markets | Unpublished | Africa | Nigeria | Lower Middle | Online | Design | J7 | Labor market | Gender discrimination |
| Ndayikeza (2025) | Underemployment of college graduates: is doing anything better than doing nothing?. | Journal of Development Economics | Africa | Burundi | Low | In person | Both | J2 | Labor market | Returns to experience |
| Bartős et al. (2024) | Gendered Access to Finance: The Roles of Team Formation, Idea Quality, and Implementation Constraints in Business Evaluations | Unpublished | Africa | Uganda | Low | In person | Both | G2 | Credit market | Gender discrimination |
| Beam et al. (2024) | Improving Gender Norms in the Workplace | Unpublished | Asia | Bangladesh | Lower Middle | In person | Implementation | J7 | Labor market | Gender norms |
| Bohren et al. (2025) | Systemic discrimination: Theory and measurement | Quarterly Journal of Economics | North America | U.S.A | High | Online | None | J7 | Labor market | Gender, race discrimination |
| Bustelo et al. (2023) | What is the price of freedom? Estimating women's willingness to pay for job schedule flexibility. | Economic Development and Cultural Change | South America | Colombia | Upper Middle | In person | Implementation | J7 | Labor market | Gendered job preferences |
| Chan (2022) | Discrimination against doctors: A field experiment | Unpublished | North America | U.S.A | High | Online | Both | I1 | Health market | Race discrimination |
| Chen (2025) | Is Deservingness Merit-based or Need-based? College Selectivity and Medical Crowdfunding Outcomes | Unpublished | Asia | China | Upper Middle | Online | None | I2 | Charitable giving | Deservingness |
| Colonnelli et al. (2024a) | Investing with the government: A field experiment in China | Journal of Political Economy | Asia | China | Upper Middle | Online | None | G2 | Investor market | Entrepreneurship |
| Colonnelli et al. (2024b) | Do Information Frictions and Corruption Perceptions Kill Competition? A Field Experiment On Public Procurement in Uganda | Unpublished | Africa | Uganda | Low | In person | Implementation | G3 | Public procurement | Corruption |
| Colonnelli et al. (2022) | Politics at work | American Economic Review | South America | Brazil | Upper Middle | Online | Implementation | J7 | Labor market | Political preferences |
| Colonnelli et al. (2025b) | ESG Is the Most Polarizing Nonwage Amenity: Evidence from a Field Experiment in Brazil | AEA: Paper & Proceedings | South America | Brazil | Upper Middle | Online | Implementation | J2 | Labor market | ESG practices |
| Del Carpio and Fujiwara (2023) | Do gender-neutral job ads promote diversity? Experimental evidence from Latin America's tech sector | Unpublished | South America | Chile | Upper Middle | Online | Both | J1 | Labor market | Gendered language |
| Ebrahimian and Zhang (2025) | Matching and Bargaining in Entrepreneurial Finance: Evidence From an Experimental System | Unpublished | North America | U.S.A | High | Online | None | G2 | Investor market | Bargaining |
| Erda and Shrader (2023) | Appearance Norms in Academic Settings. | No Working Paper | North America | U.S.A | High | Online | Design | I2 | Academeia | Appearance effects |
| Exley et al. (2024) | Information-optional policies and the gender concealment gap | Unpublished | North America | U.S.A | High | Online | None | J1 | Labor market | Gender concealment gap |
| Feng et al. (2024) | Statistical Discrimination in Two-sided Matching Markets: Experimental and Theoretical Evidence | Unpublished | North America | U.S.A | High | Online | None | G2 | Investor market | Gender discrimination |
| Ferlenga (2025) | Symbols of Oppression: The Role of Confederate Monuments in the Great Migration | Unpublished | North America | U.S.A | High | Online | Both | R2/J1 | Housing market | Location decisions |
| Fiorin et al. (2022) | Labor Markets Matching in Indonesia: Estimating Employers' Preferences with Incentivized Resume Rating | No Working Paper | Asia | Indonesia | Lower Middle | Online | None | J2 | Labor market | Employer preferences |
| Gallen and Wasserman (2023) | Does information affect homophily? | Journal of Public Economics | North America | U.S.A | High | Online | Both | J1 | Education | Gendered preferences |
| Gentile et al. (2023) | Barriers to entry: Decomposing the gender gap in job search in Urban Pakistan | Unpublished | Asia | Pakistan | Lower Middle | In person | Both | J2 | Labor market | Gender discrimination |
| Sun and Huo (2020) | The Return to Foreign and Domestic Master's Degrees in the Chinese Labor Market | Unpublished | Asia | China | Upper Middle | In person | Both | J2 | Labor market | Returns to skills/education |
| Kessler et al. (2019) | Incentivized Resume Rating: Eliciting Employer Preferences without Deception | American Economic Review | North America | U.S.A | High | Online | None | J7 | Labor market | Gender discrimination |
| Low (2024) | Pricing the biological clock: The marriage market costs of aging to women | Journal of Labor Economics | North America | U.S.A | High | Online | Implementation | J1 | Dating market | Biological clock |
| Macchi and Raisaro (2025) | Hidden Gender Discrimination | Unpublished | Africa | Uganda | Low | In person | Both | J7 | Labor market | Gender discrimination |
| Macchi (2023) | Worth your weight: experimental evidence on the benefits of obesity in low-income countries | American Economic Review | Africa | Uganda | Low | In person | Both | G2 | Credit market | Weight discrimination |
| Martin (2024) | Women Seeking Jobs with Limited Information: Evidence from Iraq | Unpublished | Asia | Iraq | Upper Middle | Online | Both | J7 | Labor market | Gendered preferences |
| Nielsen and Rigotti (2022) | Revealed Incomplete Preferences | Unpublished | North America | U.S.A | High | Online | Design | D9 | Lotteries | Incomplete preferences |
| Park (2025) | Patterns of Discrimination and the Effect on Model Minorities | Unpublished | North America | U.S.A | High | In person | Both | J7 | Labor market | Race discrimination |
| Raisaro (2023) | Incentives Justifying Nonconformity: Experimental Evidence from Motortaxi Organizations in Uganda | Unpublished | Africa | Uganda | Low | In person | Both | D9 | Labor market | Conformity |
| Spadavecchia et al. (2023) | Managerial implicit stereotypes and where to find them: Evidence from Incentivized Resume Rating | No Working Paper | Europe | Italy | High | Online | None | J7 | Managers | Gender discrimination |
| Low et al. (2023) | Estimating Employer Preferences Over Workers in India | No Working Paper | Asia | India | Lower Middle | Online | Design | J2 | Labor market | Employer preferences |
| Zhang and Zou (2025) | ESG Aversion: Experimental Evidence on Perceptions and Preferences | Unpublished | North America | U.S.A | High | Online | Design | G2 | Investor market | ESG preferences |
| Zhang (2024) | Discrimination in the Venture Capital Industry: Evidence from Two Randomized Control Trials | Management Science | North America | U.S.A | High | Online | Design | G2 | Investor market | Gender, race, age discrimination |
| Sinha and Zhu (2020) | Labor Market Returns to Upskilling - A Combination of Audit Study and Resume Review | No Working Paper | North America | U.S.A | High | Online | None | J2 | Labor market | Returns to skills/education |

**Notes:** This table lists all papers included in our review of published and unpublished papers that leverage the IRR approach. Unpublished includes working papers or pre-registrations. when doesn't sume to 1.

# A Appendix

## A.1 Classification Process

This appendix describes the criteria used to classify papers in our literature review of incentivized ratings design (IRR) studies.

**In-Person vs. Online:** Studies were classified as in-person if decision makers completed evaluations during face-to-face meetings with research staff or enumerators. This includes cases where enumerators visited workplaces to administer the task, decision makers came to a central location (e.g., research office, training venue) to participate, or if decision makers were recruited in person. Studies were classified as online if decision makers completed evaluations remotely via web-based platforms (e.g., Qualtrics) without direct researcher supervision during the task. In cases where implementation mode varied within a study (e.g., some participants online, some in person), we coded the study according to the predominant mode.

**Design Twists:** A study was classified as having a design twist if it departed from the original Kessler, Low, and Sullivan (2019) paradigm by: (1) modifying how applications are created (applications duplicated and names/gender/photo randomized); (2) using alternative incentive mechanisms (e.g., payments based on candidate outcomes, providing advice rather than referrals); (3) implementing information treatments to test discrimination mechanisms (e.g., randomizing quality signals); or (4) embedding additional experimental interventions (e.g., training decision makers, two-stage designs).

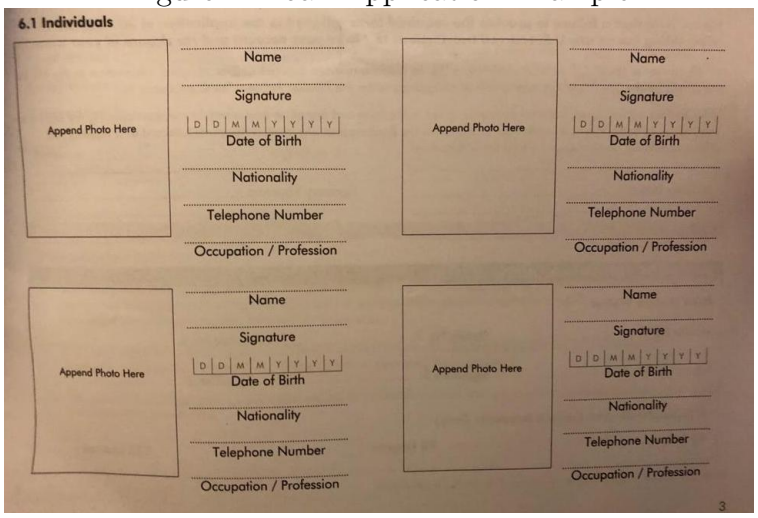
**Implementation Twists:** A study was classified as having an implementation twist if it modified the mechanics of profile presentation or evaluation by: (1) using conjoint or choice-based formats (e.g., choosing within a menu, pairwise comparisons) rather than rating profiles independently; (2) employing non-standard randomization (e.g., pre-generated shells with

within-profile randomization, using real CVs with selective manipulation); (3) using an in person design

**Setting and Topic** Each study was coded for its primary empirical setting (e.g., labor market, credit market, academia, investment, dating) and main research topic (e.g., gender discrimination, racial discrimination, age discrimination, preference elicitation, statistical discrimination, hiring frictions). Several common topics and settings, such as labor market and gender, were also separately encoded in their own categories. These categories were not mutually exclusive, and some studies were assigned multiple topic codes when they addressed several research questions.

## A.2 Appendix Figures

Figure 4: Loan Application Example




*Note:* Loan application form example from Macchi (2023).