The Judgment Bias Task: A flexible method for assessing individual differences in social judgment biases

Jordan R. Axt[a,*], Helen Nguyen[a], Brian A. Nosek[a,b]

[a] University of Virginia, United States
[b] Center for Open Science, United States

ARTICLE INFO

ABSTRACT

Many areas of social psychological research investigate how social information may bias judgment. However, most measures of social judgment biases are (1) low in reliability because they use a single response, (2) not indicative of individual differences in bias because they use between-subjects designs, (3) inflexible because they are designed for a particular domain, and (4) ambiguous about magnitude of bias because there is no objectively correct answer. We developed a measure of social judgment bias, the Judgment Bias Task, in which participants judge profiles varying in quality for a certain outcome based on objective criteria. The presence of ostensibly irrelevant social information provides opportunity to assess the extent to which social biases undermine the use of objective criteria in judgment. The JBT facilitates measurement of social judgment biases by (1) using multiple responses, (2) indicating individual differences by using within-subject designs, (3) being adaptable for assessing a variety of judgments, (4) identifying an objective magnitude of bias, and (5) taking 6 min to complete on average. In nine pre-registered studies ($N > 9000$) we use the JBT to reveal two prominent social judgment biases: favoritism towards more physically attractive people and towards members of one's ingroup. We observe that the JBT can reveal social biases, and that these sometimes occur even when the participant did not intend or believe they showed biased judgment. A flexible, objective, efficient assessment of social judgment biases will accelerate theoretical and empirical progress.

1. Introduction

Social bias – intended or unintended favoritism in evaluation, judgment, or behavior for one social group over another – is pervasive. Sometimes people are aware of their biases and embrace them as guides for behavior. For example, the first author only watches Duke basketball games with people willing to cheer for Duke, disqualifying the second and third authors. Other times, biases differ from conscious values, and can cause actions to deviate from intended behaviors. Discrimination in hiring (Ameri et al., 2015), academic (Milkman, Akinola, & Chugh, 2012), and economic (Doleac & Stein, 2013; Edelman, Luca, & Svirsky, 2017) contexts may occur without conscious intention to discriminate, or awareness of doing so (Bertrand, Chugh, & Mullainathan, 2005; Bertrand & Duflo, 2016; Rooth, 2010).

The social consequences of biases, combined with the possibility that some occur outside of intention or awareness, have made them a popular topic of research. At the same time, there are pervasive methodological limitations for conducting controlled experimental research on judgment biases including low reliability, lack of insight on individual differences in degree of bias, lack of an objective standard indicating no bias, and idiosyncratic paradigms that cannot be adapted for multiple uses.

Low reliability. Most bias investigations rely on a single judgment or behavior as the dependent variable. In 2015, there were 68 studies testing a judgment or behavioral preference for one social group over another published in four social psychology journals: *Journal of Personality and Social Psychology*, *Personality and Social Psychology Bulletin*, *Journal of Experimental Social Psychology*, and *Psychological Science*.[1] Of them, 47 (68%) relied on only a single judgment or behavior for bias assessment, and 57 (83%) relied on five or fewer. Examples of single-shot outcomes included allocating resources (Binning, Brick, Cohen, & Sherman, 2015) or providing hypothetical prison sentences (Cheung & Heine, 2015). Single response assessments, particularly of social judgments or behaviors that are influenced by a variety of factors, are often unreliable and weaken power to detect biases. Underpowered research increases the rate of Type 1 and Type 2 errors (Button

---

* Corresponding author at: University of Virginia, Box 400400, Charlottesville, VA 22904-4400, United States.
  E-mail address: jra3ee@virginia.edu (J.R. Axt).

[1] We included all studies that used actual behavior or behavioral intentions between existing social groups. See https://osf.io/u2mbx/ for list of all included studies and outcome measures.

et al., 2013) and contributes to weakening reproducibility of research (Asendorpf et al., 2013; Funder et al., 2014).

Measuring individual differences. Many existing bias paradigms are unable to distinguish the relative strength of biased behavior between participants. Partly this is a function of lower reliability based on single responses. Another contributor is reliance on between-subjects designs. For example, in Norton, Vandello, and Darley (2004), participants chose between two fictional college applicants. Candidates had different strengths, with one applicant being Black and the other White, and race randomly assigned to strengths between subjects. Black applicants were favored regardless of condition, indicating racial bias in the aggregate. These studies were not focused on finding individual differences in social judgment bias, but there may be added benefits to developing within-subjects measures to estimate the social bias for each participant. Such a design enables assessment of group-level differences (e.g., the impact of an intervention on reducing levels of racial bias in judgment) and individual differences (e.g., the relation between racial attitudes and racial bias in judgment).

Objective standard. Many measures of bias have no objectively correct answer, meaning bias can only be understood in relative terms between participants or conditions (e.g., Blommaert, van Tubergen, & Coenders, 2012). For example, Haddock, Zanna, and Esses (1993) used a hypothetical budget paradigm to study attitudes towards gay people. Participants needed to cut funding for several organizations, one of which was the university's gay and lesbian organization. More prejudiced participants proposed harsher reductions in funding towards the gay and lesbian organization. However, there is no objective standard for what level of funding indicates lack of bias. As a consequence, there is no way to identify who is biased and to what extent they are biased.

It is often of practical, legal, and theoretical interest to know if social judgments conform to an objective standard. If measures can only represent biased behavior in relative terms, then it is not possible to investigate or conclude when a judgment or behavior is unbiased.

Adaptability for multiple uses. Implicit measures like the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998; Nosek, Greenwald, & Banaji, 2007) are used frequently, in part, because they can be adapted to a variety of domains. To measure new content, researchers retain the established procedural parameters and change just the task stimuli following established best practices (Lane, Banaji, Nosek, & Greenwald, 2007). For many social judgment bias measures, the procedure and content are not easily separated, making it difficult to adapt the method for other uses. For example, measures investigating social bias through employment resumes cannot be easily adapted to other forms of social bias. Moreover, measures like the IAT are reliable and efficient to administer by collecting multiple responses quickly, which maximizes applicability across research contexts.

Given limitations of existing measures, we sought to develop a measure of social judgment bias that (1) maximized effective reliability, (2) is sensitive to measuring well-known biases, (3) identified individual differences in bias, (4) can identify magnitude of bias compared to an objective standard, (5) is efficient to administer, and (6) is flexible for a variety of uses.

1.1. The Judgment Bias Task

Prior studies on intergroup bias used methods that share some of the intended strengths of the Judgment Bias Task (JBT). For example, some studies asked participants to predict individuals' future behavior based on profiles that included both diagnostic information and irrelevant social information (e.g., gender; Beckett & Park, 1995; Locksley, Hepburn, & Ortiz, 1982). Equating the diagnostic information across social categories enabled assessment of the impact of the social information in forming predictions. Likewise, conjoint analysis reveals social bias by asking participants to choose between multiple pairs of targets who vary on levels of both task-relevant information and task-

irrelevant social information (e.g., perceived weight; Caruso, Rahnev, & Banaji, 2009). By equating targets on task-relevant information across social groups, conjoint analysis can reveal the extent to which social information influences choices. Finally, Situational Judgment Tests (SJTs), common in personnel psychology (e.g., Cabrera & Nguyen, 2001), present participants with hypothetical and ambiguous scenarios and ask them to rank potential responses. Researchers can design SJTs to measure social judgment biases often not aware to participants.

The JBT builds on some of the features of these paradigms to assess social judgment biases. In the JBT, participants evaluate a series of profiles for a particular outcome, such as membership in an honor society or selection of team members. Each profile has multiple quantified criteria that are relevant for decision-making and one or more that are ostensibly irrelevant. Participants are instructed to weigh the relevant criteria equally in their judgment. The profiles are constructed so that some are systematically better than the others, but the difference is somewhat difficult to detect. Participants are assessed on their sensitivity to distinguishing between the better and worse profiles, and whether they have a bias to be more lenient or stringent to candidates with different irrelevant criteria.

One example of a JBT involves instructing participants to accept approximately half of the applicants to a hypothetical honor society. Each applicant profile has four pieces of relevant information: Science GPA, Humanities GPA, recommendation letter strength, and interview score. Simultaneously, ostensibly irrelevant gender information is communicated with a face accompanying the profile. Unobtrusively, a random half of the male and female profiles are made somewhat more qualified than the others. Participants then evaluate the individual profiles sequentially to make accept-reject decisions. Each participant's performance produces scores for their ability to distinguish more from less qualified applicants, and whether judgments were more lenient or strict compared to the objective standard, both overall and separately for each gender.

Unlike past work using related methods investigating intergroup bias (Beckett & Park, 1995; Cabrera & Nguyen, 2001; Locksley et al., 1982), the JBT is analyzed using Signal Detection Theory (SDT). Decisions made during the task can be assessed based on *sensitivity* (d') and *criterion* (c). Sensitivity measures the extent to which a participant distinguishes more from less qualified profiles. Participants with high sensitivity are better at accepting the more qualified and rejecting the less qualified profiles than those with low sensitivity. A score of zero indicates no ability to distinguish more from less qualified profiles.

Criterion measures the extent to which a participant is lenient or strict in evaluation. Lower criterion values indicate being more lenient, and higher criterion values indicate being more strict. A score of zero indicates equal likelihood of correctly accepting more qualified profiles and correctly rejecting less qualified profiles. By computing separate sensitivity and criterion estimates for each of the social groups in the task, the JBT measures whether participants are better able at discriminating between more and less qualified profiles and whether the criterion for acceptance differs between social groups. SDT has been used productively in implicit measures of bias such as the Go/No-Go Association Task (Nosek & Banaji, 2001) and "shooter bias" tasks (Correll, Park, Judd, & Wittenbrink, 2002).

Participants may show socially biased judgment on the JBT for a variety of reasons. For example, in a JBT assessing gender biases in academic honor society admissions, some participants may have a lower acceptance criterion for male than female applicants because they believe males are more academically gifted than females, or because they simply prefer males to females. In these cases, bias on the task is intentional. Alternatively, some participants may have a lower acceptance criterion for male than female applicants even if they wanted to treat applicants from both genders equally and believe they did so. In these cases, participants' judgments may be shaped by processes operating outside of conscious awareness or intention, such as prominent, culturally-based associations between gender and