How-To Guide for the Judgment Bias Task

Jordan Axt, Helen Nguyen & Brian Nosek

The Judgment Bias Task (JBT: Axt, Nguyen & Nosek, 2016) is a fast, flexible, and reliable paradigm to measure bias in decision making. This document is a general guide to the structure of the JBT and how it can be adapted for other research purposes.

Before starting, it may be helpful to run through a sample version of the JBT yourself. A sample Inquisit study is available for download at jordanaxt.com. You do not need to purchase an Inquisit license to complete the sample study.

In general, the JBT asks participants to evaluate a series of **profiles**. Profiles are created beforehand to have differing **levels** (e.g., better students vs. worse students, healthier patients vs. less healthy patients, etc.). Within each level, the profiles are then assigned to different social **categories** (e.g., more attractive people vs. less attractive people, people from one country vs. another country, etc.). Researchers can then see if the presence of these categories biases the way the profiles are evaluated.

For example, in Axt, Nguyen and Nosek (2016) Study 2a and Study 2b, the JBT was used to investigate own-university bias when making decisions about an academic honor society. Participants evaluated 64 **profiles** for an academic honor society. These profiles had qualifications that placed them into two **levels** (32 profiles were more qualified applicants and 32 profiles were less qualified applicants). Within these levels, profiles were then assigned to two different **categories** (applicants from the same university vs. applicants from a rival university). The study then investigated whether university affiliation impacted participants' ability to accept the more qualified and reject the less qualified profiles into the academic honor society.

Below is more specific information about how to decide on the levels, profiles and categories you'll need to the JBT. Afterwards, we go over information on how to program and analyze the JBT.

Getting Started: Deciding on the Structure of Your JBT

When using the JBT, researchers must decide ahead of time on what levels and categories are going to be used.

Levels

Levels are designed to suit the bias context you are interested in. Researchers interested in bias within a hiring context may design a JBT that asks participants to evaluate applicants for a job using levels or more qualified vs. less qualified applicants. Or, researchers interested in bias in an athletic context may design a JBT that asks participants to evaluate which players to select for an athletic team using levels of more athletically skilled vs. less athletically skilled.

Levels need to be designed such that there can be objectively correct responses. For example, participants tasked with selecting the best athletes for a team will be correct every time they accept a more athletic profile and incorrect when accepting a less athletic profile.

This decision of what context to use for the JBT is driven partly by the area that researchers may be interested in and partly by what the participants in the study may be more familiar with (e.g., undergraduates may feel more comfortable selecting applicants for a student group than selecting soldiers for a military unit).

Once researchers have decided on the levels they'd like to use on their task, the next step is to create the profiles that belong in each level.

Profiles

Profiles are created by first developing a list of qualifications, and then standardizing those qualifications to create individual profiles belonging to each level.

Below are two examples of creating a list of qualifications and then using those qualifications to produce profiles for different versions of the JBT.

Profile Examples

Example 1: An Academic Context

In Axt, Nguyen & Nosek (2016), the JBT was used to evaluate bias in an academic context. Participants were asked to evaluate applicants for an academic honor society. There were two levels in the task: more qualified applicants and less qualified applicants.

To create the two levels, profiles were shown with four qualifications: Science GPA, Humanities GPA, Letters of Recommendation, and Interview Score. The two GPA's were presented on a scale from 1-4. The recommendation letters also had four levels (poor, fair, good, or excellent). The interview score was on a 1-100 scale. Here is an example profile:

Science GPA:	3.5
Humanities GPA:	3.7
Recommendation Letters:	Excellent
Interview Score:	70

For scoring, these qualifications were all then standardized to have a 1-4 range. The GPA's were already on a 1-4 scale. The recommendation letter scores were assigned a numeric values (Poor=1, Fair=2, Good=3, Excellent =4). The interview score was divided by 25.

Using this 1-4 scoring, profiles in the more qualified level were created to have a total score of 14. Here is an example of a more qualified profile:

Science GPA:	3.8
Humanities GPA:	3.6
Recommendation Letters:	Good
Interview Score:	90

Based on the 1-4 scoring, this profile has a score of 3.8 + 3.6 + (Good=3) + (90/25=3.6) = 14. For the more qualified level, we then made 31 additional, unique profiles with qualifications that summed to 14.

For the less qualified level, profiles were created to have a total score of 13. Here is an example of a less qualified profile:

Science GPA:	3.9
Humanities GPA:	3.2
Recommendation Letters:	Good
Interview Score:	72.5

Based on the 1-4 scoring, this profile has a score of 3.9 + 3.2 + (Good=3) + (72.5/25=2.9) = 13. For the less qualified level, we then made 31 additional, unique profiles with qualifications that summed to 14.

The task then had 64 trials, with 32 more qualified profiles and 32 less qualified profiles. Trials were scored as correct whenever participants accepted a more qualified or rejected a less qualified profile. Trials were coded as incorrect whenever participants accepted a less qualified and rejected a more qualified profile.

Example 2: Romantic Context
In another study in Axt, Nguyen & Nosek, 2016), the JBT was used to evaluate bias in a romantic context. Participants were asked to evaluate profiles for an online dating site. There were two levels in the task: more attractive profiles and less attractive profiles.

To create the two levels, profiles were shown with six qualifications: Attitude Similarity, Social Similarity, Intelligence, Openness, Dependability and Sense of Humor. The two similarity scores were presented on a 1-10 scale. The intelligence and openness scores

were presented on a 1-4 scale. The dependability and sense of humor scores had four levels (poor, fair, good, or excellent). Here is an example profile:

Attitude Similarity:	7.5
Social Similarity:	6.5
Intelligence:	3.5
Openness:	3.4
Dependability:	Good
Sense of Humor:	Excellent

For scoring, these qualifications were all then standardized to have a 1-4 range. The intelligence and openness scores were already on a 1-4 scale. The dependability and sense of humor scores were assigned a numeric values (Poor=1, Fair=2, Good=3, Excellent =4). The attitude and social similarity scores were divided by 2.5.

Using this 1-4 scoring, profiles in the more attractive level were created to have a total score of 21. Here is an example of a more attractive profile:

Attitude Similarity:	9.5
Social Similarity:	8.5
Intelligence:	3.6
Openness:	3.2
Dependability:	Good
Sense of Humor:	Excellent

Based on the 1-4 scoring, this profile has a score of (9.5/2.5=3.8) + (8.5/2.5 = 3.4) + 3.6 + 3.2 + (Good=3) + (Excellent=4) = 21. For the less attractive level, we then made 29 additional, unique profiles with qualifications that summed to 21.

For the less attractive level, profiles were created to have a total score of 19.5. Here is an example of a less attractive profile:

Attitude Similarity:	9.0
Social Similarity:	7.5
Intelligence:	3.2
Openness:	2.7
Dependability:	Good
Sense of Humor:	Excellent

Based on the 1-4 scoring, this profile has a score of (9.0/2.5=3.6) + (7.5/2.5 = 3) + 3.2 + 3.7 + (Good=3) + (Excellent=4) = 19.5. For the less attractive level, we then made 29 additional, unique profiles with qualifications that summed to 19.5.

The task then had 60 trials, with 30 more attractive profiles and 30 less attractive profiles. Trials were scored as correct whenever participants accepted a more attractive or rejected a less attractive profile. Trials were coded as incorrect whenever participants accepted a less attractive and rejected a more attractive profile.

Making Your Own Profiles

Selecting Qualifications

Qualifications should be selected so that they appear as objective as possible. In general, it is best to select qualifications that are not all on the same scale. This will prevent participants from simply summing the qualifications for each profile and then using that sum to guide their decision-making. Similarly, it's best to select more than two or three qualifications for each profile. More qualifications will make it harder for participants to detect the different levels being used in the JBT.

Scoring Profiles

There are no guidelines for how to score profiles to assign them to different levels. Scoring the levels too closely to one another (for example, more qualified applicants having a score of 14 and less qualified applicants having a score of 13.9) will decrease overall accuracy and may result in chance levels of performance.

On the other hand, scoring the levels too far apart (for example, more qualified applicants having a score of 14 and less qualified applicants having a score of 7) will increase overall accuracy and likely result in less bias on the task. Researchers should strive for scoring that creates accuracy levels that are above chance but still far from ceiling.

Number of Profiles

There is no set number of profiles to have within each level, but more profiles in each level increases the reliability of the JBT and decreases the ability for participants to monitor

their potential biases. Axt, Nugyen & Nosek (2016) used a minimum of 20 profiles for each level.

Once you have identified the levels you'll be using and made the profiles for each level, the next step is to identify the categories that will be used in the JBT.

Categories

Categories are selected based on the social bias researchers are interested in studying. Selecting categories is likely going to be the most straightforward selection made when using the JBT. For instance, researchers interested in studying bias towards Whites vs. Hispanics would have one category be White people and another category be Hispanic people. Researchers interested in studying bias towards thin vs. obese people would have one category be thin people and another category be obese people.

In the JBT, stimuli from these categories will then be paired with profiles from both levels. For example, in the romantic context example above using White and Hispanic people as categories, half of the more attractive profiles would be paired with White people and half would be paired with Hispanic people. Likewise, half of the less attractive photos would be paired with White people and half with Hispanic people.

Category stimuli should be selected such that there are as many items from each category as there are profiles in each level. For example, a JBT with 60 trials and 30 profiles in each level should also have 30 stimuli from each category.

Programming the JBT

A sample Inquisit program for the JBT is available at jordanaxt.com.

This program is designed so that you will only need to change the images you are using for your profiles and categories. You will also need to update the task instructions to fit your specific use of the JBT. The program is annotated throughout to tell you what each component is doing and what specific portions of the code you'll need to change in order to run your own JBT study.

Analyzing the JBT

The JBT is primarily analyzed using Signal Detection Theory (SDT). Through an SDT analysis, decisions made during the task can be assessed based on their *sensitivity* (d') and their *criterion* (c). These criterion and sensitivity measures are computed separately for each category used in the JBT.

Sensitivity measures the extent to which a participant is able to distinguish the two levels of profiles. For example, participants with high sensitivity for one category are better at accepting the more qualified profiles and rejecting the less qualified profiles than those with low sensitivity for that category.

Criterion is the decision threshold. In the context of the JBT, criterion measures the level at which any profile below the threshold is rejected and any profile above that threshold is accepted. Lower criterion values for a category would indicate being more lenient, and higher criterion values would indicate being more restrictive for that category.

By computing separate sensitivity and criterion estimates for each of the categories used in the task, the JBT measures whether participants are better able at discriminating between the levels of profiles in one category over another as well as whether the criterion for acceptance differs between categories.

SPSS and R cleaning scripts for the JBT are available at jordanaxt.com. These cleaning scripts will calculate the following for each participant:

1) Overall task accuracy
2) Accuracy on trials with Category A
3) Accuracy on trials with Category B
4) Overall rate of providing one response (for example, useful for screening out participants who accepted way too many or way too few profiles)
5) Whether participant provided one response to all profiles from Category A (for example, useful for screening out participants who accepted or rejected all applicants from a specific category).
6) Whether participant provided one response to all profiles from Category B
7) Sensitivity for Category A
8) Sensitivity for Category B
9) Criterion for Category A
10) Criterion for Category B

Using More Than Two Categories

The examples in this document have focused on JBT's using only two categories. However, the JBT needs only *at least* two categories; more categories can be added depending on researcher needs. For example, in Axt, Nguyen & Nosek (2016), racial bias in a romantic context was investigated using two levels (more attractive vs. less attractive profiles) but three racial categories (White, Black and Hispanic people). There were 60 total trials, with 30 more attractive profiles and 30 less attractive profiles. Within each level of profile, 10 were White, 10 were Black, and 10 were Hispanic. The Inquisit programs and analysis scripts provided can be adapted to have more than two categories.

Lingering Questions

If you have any remaining questions about how to use the task, please contact Jordan Axt (jaxt@virginia.edu).