Characterizing Counterstereotypes in terms of Bayesian Probabilities

Arkesh Das

Department of Psychology, Michigan State University

PSY 235: Social Psychology

Prof. Cesario

12/08/2023

Abstract

Recent research has been conducted to bolster the findings of the classical study done by McCauley and Stitt on the Bayesian nature of stereotypes. These sets of experiments were conducted on a variety of traits and social groups. This study verifies that individuals use Bayesian probabilities to model stereotypes across various social groups. Additionally, it once again demonstrated that stereotypical traits can be identified from their diagnostic ratios. Stereotypic traits for each group generally had diagnostic ratios greater than one. Additional traits analyzed for each group not considered stereotypes were labeled “non-stereotypes". These traits generally had diagnostic ratios less than one. However, among the traits considered to be non-stereotypes, certain traits had diagnostic ratios much smaller than the diagnostic ratios of other non-stereotypic traits. These traits tended to be traits that were not only “non-stereotypic”, but could also be considered “counterstereotypic” to a group, meaning they could be considered traits that characterize a group much less than the general population. This indicates that there may be a distinction between the diagnostic ratios of non-stereotypic and counterstereotypic traits. Further data must be collected to explore this relationship. The goal of this proposal is to describe 1) a way to determine the presence of counterstereotypic traits and non-stereotypic traits for certain groups, 2) a method to determine if these counterstereotypic traits display a significant difference in their diagnostic ratios when compared to non-stereotypic traits and 3) a method to finding the “cut-off” between the diagnostic ratios of non-stereotypic and counterstereotypic traits.

Characterizing Counterstereotypes in terms of Bayesian Probabilities

Stereotypes are an integral tool that we as humans use as a part of our process of analyzing the world around us. The initial definition of stereotypes in the 1920s defined them as being inherently "factually incorrect" (Lippman 1922). However since then, a more nuanced definition of stereotypes emerged, one that has radically altered the way stereotypes are viewed in the field. Later studies have shown that stereotypes can be accurate, especially racial and gender based stereotypes (Jussim et al. 2009).

Beginning in the early to mid 20th century, attempts were made towards quantifying stereotypes. Initial analyses used survey responses to qualitatively rank traits in terms of how much they characterized a group (Katz and Braly 1933). However, in a now classical paper, McCauley and Stitt asserted that stereotypes and the act of stereotyping could be modeled in terms of Bayesian posterior probabilities (McCauley and Stitt 1978). This claim was derived from the fact that humans use Bayes' theorem when making predictions, and that stereotypes are treated as a type of prediction by the human mind: one that is largely based on previously held beliefs, be it those derived from first-hand experiences or shared cultural contexts.

Formula and Premise of Bayes' Theorem

The basic formula for Bayes' theorem is as follows, where A and B are two non-mutually exclusive and dependent events:

$$P(A|B) = \frac{P(A)*P(B|A)}{P(B)} \quad (1)$$

Therefore, given previous information on the probability of A occurring ($P(A)$), the probability of B occurring($P(B)$), and the probability of B given A ($P(B|A)$), the probability of A and B can be calculated ($P(A|B)$).

Application of Bayes' Theorem to Predictive Stereotyping

This general formula can be adapted to the context of stereotypic predictions in the following way, where $trait$ represents a certain measurable aspect of an individual and $group$ is the social group or "category" that an individual is a part of:

$$P(trait|group) = \frac{P(trait)*P(group|trait)}{P(group)} \quad (2)$$

In this context, $P(A|B)$ represents the probability of an individual having a specific trait ($P(trait|group)$). This is called the posterior probability, or the probability after taking into account $P(trait)$, $P(group)$, and $P(group|trait)$. $P(trait)$ is $P(A)$, which in this context is the probability of an individual in the general population having a specific trait. $P(group)$ corresponds to $P(B)$ and represents the probability of an individual in the general population belonging to a social category. Finally, $P(group|trait)$ is the probability of an individual being part of a specific social category given that they have a specific trait.

Why might people use Bayes' Theorem?

Social predictions are important in gaining information about individuals when no individuating information is available, for example when meeting a person for the first time. In the presence of individuating information, such as in the form of a description, the predictions that people make do not follow Bayes' Theorem (Kahneman and Tversky 1973). However, Bayes' Theorem is the most mathematically optimal way to predict what the frequency of a trait in a group is, given prior knowledge of both the trait and the general population and no additional information about the individual. Logically, it makes sense that the social categories that an individual belongs to influences the likelihood of them having certain traits, and that the more that a trait is associated with a group, the higher probability you can attribute to an individual having a specific trait given that they belong to a group.

The Diagnostic Ratio

An additional calculation can be derived from Bayes' Theorem, that of the likelihood ratio, or the diagnostic ratio. Applied to context of social predictions, the diagnostic ratio (DR) can be written as:

$$DR = \frac{P(trait|group)}{P(trait)} \quad (3)$$

The diagnostic ratio is an important calculation since it tells us the "strength" of any given prediction. If the diagnostic ratio is large, then that means that the knowledge that an individual is part of a certain social category greatly should increase the probability that an individual has a specific trait.

A Recap on McCauley and Stitt

To test this idea that humans follow the logic of Bayes' Theorem under conditions of little to no individuating information, McCauley and Stitt conducted a series of questionnaires in which Ps were asked to provide estimations for the probabilities corresponding to the four components of Bayes' Theorem (Equation 2), as shown below:

1) What percentage of all the world's people are (trait)?
2) What percentage of all the world's people are (trait)?
3) What percentage of (group) are (trait)?
4) What percentage of all the world's people are (group)?

In the context of their first two experiments, the social category that McCauley and Stitt used was Germans. They chose to look at nine traits, four which were previously identified as stereotypes for Germans (Karlins et al 1969) and five other traits that were not listed as stereotypes for Germans, Using the responses from questions (1), (2), and (4), McCauley and Stitt calculated the expected response for question (3) using Bayes' Theorem and compared it to

the Ps actual responses to question (3). Throughout the three experiments in which these questionnaires were conducted, the order in which these questions were asked were varied, but one thing remained the same. The Ps actual probabilities for question (3) were highly correlated with their calculated responses to (3) (r =0.91) (McCauley and Stitt 1978).

However, perhaps the most interesting finding from their experiments was that the diagnostic ratios (Equation 3) between stereotypic and "non-stereotypic" traits (those that were not defined as stereotypes) were significantly different. The stereotypic traits all had diagnostic ratios that were greater than one and the non-stereotypic traits had diagnostic ratios that were diagnostic ratios that were all less than one. Therefore from these findings, McCauley and Stitt proposed a new empirical definition for stereotypes. They defined stereotypes as being traits that had diagnostic ratios greater than one for a given group.

McCauley and Stitt Revisited: Solanki and Cesario

There were quite a few limitations to McCauley and Stitts' findings. For one, their first two experiments only tested traits with respect to one social category: the national category Germans. Their third study tested stereotypes associated with black Americans, but on a very small scale. Their sample sizes for each study were 69, 52 and 75 Ps respectively (McCauley and Stitt 1978). Secondly, they ran their questionnaires on very limited populations. The Ps for studies 1 and 2 were all undergraduate women, and while study 3 had a more diverse set of Ps, there was still very limited diversity, as only four of the Ps were black.

Therefore, in order to extend the findings made by McCauley and Stitt, a much larger set of studies needed to be done. Recently, Solanki and Cesario have run three sets of experiments that not only analyzed traits for multiple social categories, but also ran these tests across a variety of populations (Solanki and Cesario 2021).

Social Categories*.** The eight social groups were chosen to represent four different 'types' of social categories: gender, race, sexuality, and profession. The two gender categories were male and female, the two racial categories were asian and black, the two sexuality categories were gay (gay male) and lesbian, and the two professional categories were lawyer and politician. In Experiment 1, ten traits for each category were compiled based on previous literature and "prior beliefs." Ps in experiment 1 were given an opportunity to list traits they believed were stereotypical for the social category they were assigned (see ***Methodology for more details), and these responses were taken into account to modify the traits presented in Experiment 2. Experiment 2 only used the top three stereotypic traits obtained from the free-response data, and three non-stereotypic traits for each of the eight social categories. Experiment 3 also incorporated those same sets of traits and categories.

***Sampling*.** The first two experiments were conducted on undergraduate students, who were enrolled in Michigan State University. This allowed for a larger range of Ps to be tested when compared to McCauley and Stitt's work. The Ps for experiments 1 and 2 were recruited through an online system (SONA) in which they could earn "credit hours" for their participation in studies. Approximately 70% of the Ps were female, 62% were white and the mean age was 19.72 years. Experiment 3 was done using Amazon's "crowdsourcing marketplace" MTurk (*Introduction to Amazon Mechanical Turk* 2018), which recruited more diverse Ps. Ps were given $0.50 for participating in the study.

***Methodology*.** Experiment 1 largely uses the same methodology as McCauley and Stitt's questionnaire, Ps were given the opportunity to provide responses for each component of the Bayesian formula, and their responses were analyzed in the same manner as in McCauley and

Stitt's work. However, the surveys were taken online instead of on paper, and each probability was paired with a graphic as a visual aid for Ps as they considered their responses.

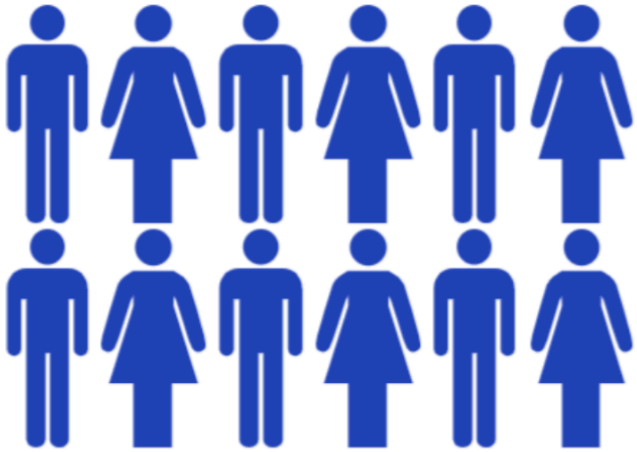

Figure 1: *Section of the Icon Array Presented to Ps* (Solanki and Cesario 2021)

Ps were only presented with the ten traits (as identified in the ***Social Categories*** section) associated with one of the eight social categories. The order in which these traits were presented was randomized. Additionally, at the end of the survey, Ps were asked about traits that they considered to be stereotypes about a different social category (not the one they answered the probability questions for). The category they were assigned to give responses for was also randomized. The question was phrased in the following way:

> Note that we are NOT asking you of your personal opinion or whether you believe these stereotypes to be true. Instead, we just want you to list the stereotypes that other people might believe about [social category]. You can list as many or as few as you would like. In the text box below, please list common stereotypes about [social category].
>
> (Solanki and Cesario 2021)

The question was written in this way to try to minimize the influence of the personal opinions that Ps had about stereotypes.

Both of the following experiments were conducted in a similar fashion, however given the smaller number of traits for each social category, additional tests were conducted to measure if certain factors correlated with the extent to which individuals' reported posterior probabilities corresponded to the calculated posterior probabilities given their responses to the other components of Equation 2. These additional tests included the use of Raven's Progressive Matrices to measure cognitive ability, as well as additional free response questions to assess political orientation and motivation to control prejudice (MCP).

Results. In general, all the results of three experiments were consistent with the findings of McCauley and Stitt. Ps generally followed Bayes' Theorem in their responses, and the diagnostic ratios for many (interestingly, not all, see **Discussion** section) stereotypic traits had diagnostic ratios greater than one, and many diagnostic ratios for non-stereotypic traits were less than one.

Potential Distinctions Among the Non-stereotypes

When taking a closer look at the diagnostic ratios of the traits that were considered non-stereotypic, an interesting trend seems to emerge. Certain non-stereotypes naturally tend to have diagnostic ratios that are much smaller than the diagnostic ratios of other non stereotypes.

In McCauley and Stitt. For example, looking at the data from the first experiment of the McCauley and Stitt study, the trait *impulsive* had a diagnostic ratio of .79 while *pleasure-loving* and *tradition-loving* have diagnostic ratios of .89 and .91 respectively (Figure 2).

Description	Judged p (trait)	Judged p (German/ trait)	Judged p (trait/ German)	Calculated p (trait/ German)	Diagnostic ratio
Efficient[a]	49.8	22.5	63.4	62.6	1.27
Extremely nationalistic[a]	35.4	23.6	56.3	46.7	1.59
Ignorant	34.0	11.9	29.2	22.6	.66
Impulsive	51.7	16.9	41.1	48.8	.79
Industrious[a]	59.8	30.4	68.2	101.6	1.14
Pleasure-loving	82.2	23.5	72.8	107.9	.89
Scientifically minded[a]	32.6	25.0	43.1	45.5	1.32
Superstitious	42.1	11.4	30.4	26.8	.72
Tradition-loving	62.4	22.2	57.2	77.4	.91

[a] Traits from German stereotype.

Figure 2: *Table 1; Mean Probabilities in Percentages and Mean Diagnostic Ratios* (McCauley and Stitt 1978)

In Solanki and Cesario. Similarly in experiment 1 of the Solanki and Cesario study this difference in diagnostic ratios is present in the non-stereotypes of many of the social categories tested. In the 'Gay' social category, *messy* (*DR* ~.6) has a much lower diagnostic ratio than *careful* (*DR* ~1.0), *studious* (*DR* ~1.1), or *ambitious* (*DR* ~1.2). Similarly, in the Lawyer category, *spiritual* (*DR* ~.6) and *emotional* (*DR* ~.5) have much lower diagnostic ratios than *caring* (*DR* ~.85)or *glamorous* (*DR* ~1.0).

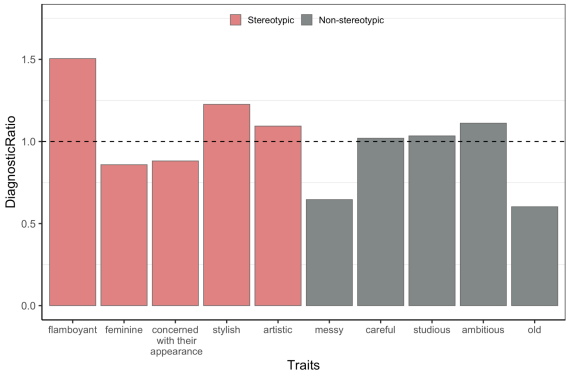

Figure 3: *Bar Graph for Diagnostic Ratio of Traits Tested in Experiment 1 for Gay Category* (Solanki and Cesario 2021)

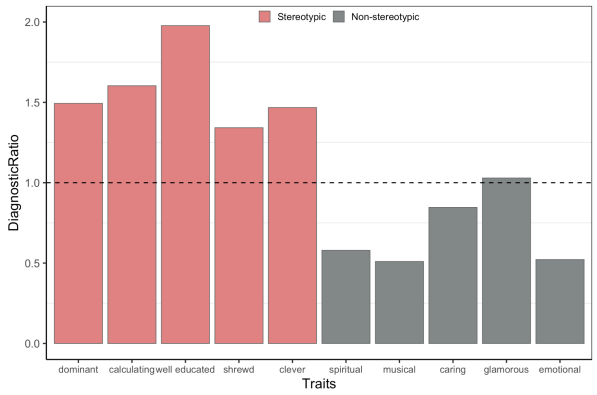

Figure 4: *Bar Graph for Diagnostic Ratio of Traits Tested in Experiment 1 for Lawyer Category* (Solanki and Cesario 2021)

***The Common Link*.** In all of these examples, the non-stereotypic trait(s) seem to not only be non-stereotypic, but could also be considered antonyms to the stereotypic traits provided for the given social category. I*mpulsive* can be seen as a near-opposite of *scientifically minded* (DR = 1.32) (Figure 2), *messy* can be considered to be a near opposite of *stylish* ($DR \sim 1.25$) , and perhaps most notably *spiritual* and *emotional* are opposites to *calculating* ($DR \sim 1.6$). In a sense, these traits could not only be considered non-stereotypes, but also "counterstereotypes".

Counterstereotypes and Their Current Context in the Literature

The idea of counterstereotypes is nothing new. For decades a small portion of literature has been dedicated to examining counterstereotypes. For example, multiple studies have been conducted on the effectiveness (or rather ineffectiveness) of 'counterstereotype interventions,' wherein Ps are shown examples of individuals who possess traits that are opposite to the stereotypic traits of the social groups that that the individual belongs to in an attempt to reduce or "suppress" the stereotype's accessibility in the mind of the Ps (Moskowitz et al 1999, Galinski and Moskowitz 2007, Burns et al 2017). However, much like stereotypes in the early 20th

century, there have been few attempts at rigorously or empirically defining what constitutes a counterstereotype.

Proposed Definitions for Counterstereotypes

Purely qualitatively, I suspect that counterstereotypes can be described simply as antonyms to traits that are considered to be stereotypic for a given social category. However, from an empirical perspective, I believe that counterstereotypes are traits that have diagnostic ratios that are *significantly* less than 1. Put in terms of probabilities, a counterstereotype are traits that are *significantly* less likely to describe a group when compared to the general population.

Overall Research Proposal

Purpose of this Research Proposal

The goal of this proposal is multifactorial. First, I aim to describe a method for qualitatively determining and distinguishing between the stereotypic, counterstereotypic and non-stereotypic traits for some given social groups. Secondly, I aim to describe an experimental procedure to determine whether counterstereotypic traits display a *significant* difference in their diagnostic ratios when compared to non-stereotypic traits of the same social group, and to determine where the "cut-off" between the diagnostic ratios of non-stereotypic and counterstereotypic traits may be. I plan on accomplishing this by conducting a study with two phases, a pre-survey and a follow-up experiment using the data collected from the pre-survey. This initial study is to serve as a "pilot" to (hopefully) more extensive future research on this topic.

Social Categories to be Tested

Given that this is a preliminary study, I will be limiting myself to analyzing only two social categories. I will be using the professional categories that were used by Solanki and

Cesario: ‘Lawyer’ and ‘Politician.’ I will explain my reason for doing so in the **Sampling and Procedure** section of **Phase I: Pre-survey**.

Phase I: Pre-survey

Purpose

The purpose of this pre-survey is to collect traits that can be used as stereotypes, counterstereotypes and non-stereotypes in the second phase of this study. The general goal of this pre-survey is to avoid having to make arbitrary guesses on what the Ps of the experiment will consider to be stereotypes, counterstereotypes and/or non-stereotypes to hopefully reduce any errors in the experiment that may result from miscategorizing a trait. I am specifically interested in collecting data on what traits that are considered to be counterstereotypic for the two social categories that I am testing.

Methodology

Similar to the Solanki and Cesario study, I plan to conduct this pre-survey over the HPR/SONA system, where Ps will earn “credit hours” for their participation in the pre-survey. I will be collecting both qualitative and quantitative data (qualitative: the responses that the Ps submit as stereotypes, etc.; quantitative: the frequency of responses). A between-subjects condition will be put in place.

Sampling, Materials and Procedure

Sampling. I suspect that the demographics of the Ps of the pre-survey will be very similar to the demographics from experiments 1 and 2 of the Solanki and Cesario study (70% female, 62% white, mean 19.72 years) seeing as they occurred only two years prior on the same campus.

Materials. Since the pre-survey will be conducted entirely online, I will only have to set up and authorize the pre-survey through SONA in order to run it.

Procedure. Consent will be gained prior to the pre-survey, where the Ps will be informed that their responses are anonymous and that they will be asked a series of questions that they should answer to the best of their abilities. The questions that will have been slightly modified from the free-response section of Experiment 1 in the Solanki and Cesario Study:

> **Q1:** Note that we are NOT asking you of your personal opinion or what you believe to be true. Instead, we just want you to list the stereotypes that other undergraduates on campus might believe about [Lawyers/Politicians]. You can list as many or as few as you would like. In the text box below, please list common stereotypes about [Lawyers/Politicians].
>
> **Q2:** Note that we are NOT asking you of your personal opinion or what you believe to be true. Instead, we just want you to list traits that other undergraduates on campus might NOT believe about [Lawyers/Politicians]. You can list as many or as few as you would like. In the text box below, please list traits that you would not expect a [Lawyers/Politicians] to have.
>
> **Q3:** Note that we are NOT asking you of your personal opinion or what you believe to be true. Instead, we just want you to list traits that other undergraduates on campus might not associate with any social group. You can list as many or as few as you would like. In the text box below, please list traits that you would not expect to be more or less likely for Lawyers, Politicians or the average person. A possible example could be *hungry*, being a Lawyer or Politician probably doesn't make you more or less likely to be *hungry* compared to the average person.

The goal of Q1 is to collect the stereotypic traits for each social category, Q2 is meant to collect the counterstereotypes for each social category, and Q3 is meant to collect non-stereotypes for

each social category. I felt that Q3 may be confusing, so I decided to provide an example for what a "possible non-stereotype" might be. Ps will be given the questions in one of two orders (between-subjects condition) as shown by Figure 5.

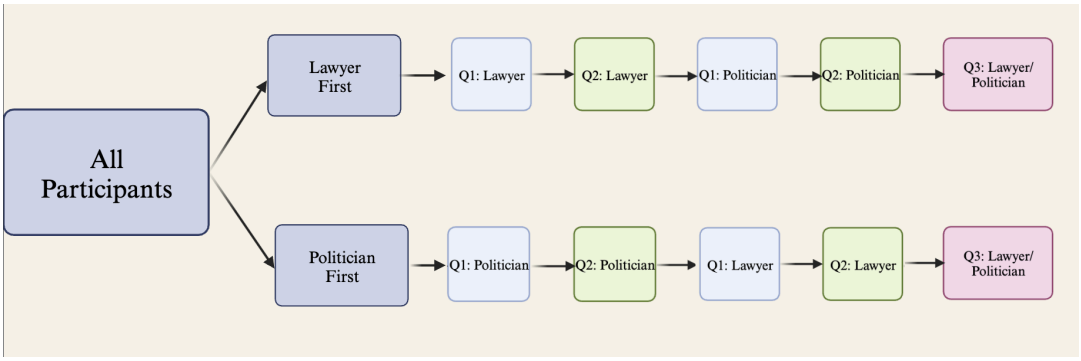

Figure 5: *Order in which Ps are asked Questions in the Pre-survey*

The reason that I am choosing to focus on professional categories (Lawyer and Politician) is because I would assume that Ps are more likely to accurately report the stereotypes, non-stereotypes and counterstereotypes for certain professions, as opposed to those for certain races, genders or sexualities.

Measurements. The primary form of measurement will be the five short answer free response questions described in the ***Procedure*** section. Subjects will be typing words or phrases that they feel best answer the questions being asked. The frequency and types of responses will be collected. The data will also most likely have to be cleaned to account for spelling errors, case sensitivity or syntax differences (e.g if someone chooses to report traits in full sentences as opposed to just a list of adjectives).

Results

Once all the pre-survey data is collected and cleaned, I will select the top three most frequent stereotypes and counterstereotypes for either category. I also plan to select three non-stereotypes for each category, but I am unsure if the frequency of the traits provided

responses will correlate to how "good" of a non-stereotype a trait is. Therefore, I will most likely have to pick three non-stereotypes arbitrarily from the responses. This means I will have nine traits for each social category.

Follow-up Experiment

Purpose

The purpose of this experiment is to determine if there is a significant difference in the diagnostic ratios of counterstereotypes and non-stereotypes and to determine what that significant difference is. Technically, I am also seeing if there is a significant difference between the diagnostic ratios of counterstereotypes and stereotypes as well.

Methodology, Sampling, Materials and Procedure

Methodology. The methodology will be almost identical to that of Experiment 1 of the Solanki and Cesario study. Instead of testing ten traits, I will be testing the nine traits identified from the **Results** from the **Pre-Survey**. There will be a between-subject to decide which social category each Ps is presented with.

Sampling. The demographics of the Ps of the experiment will be very similar to the demographics from the pre-survey.

Materials. I will most likely run this experiment online through SONA as well.

Procedure. Half of the Ps will be asked to provide their response for the four Bayesian probabilities for nine traits for the Lawyer category. The other half of the Ps will be asked for responses for the nine traits for the Politician category. Like the Solanki and Cesario study, the order in which the traits are presented to Ps will be randomized, and the order in which they are asked for the four components of the Bayes formula will also be randomized.

Results

Once the probabilities for all the traits are collected, I will then calculate the diagnostic ratios and the calculated posterior probabilities for each trait. I will be calculating the calculated posterior probabilities mainly just to verify that the data from this experiment are consistent with the results of McCauley and Stitt and Solanki and Cesario, where the calculated posterior probabilities roughly match with the actual posterior probabilities. If necessary, I will clean out any outliers.

Analysis

Now that I have the diagnostic ratios corresponding to all the traits, I can now see how they relate. First, I will plot the diagnostic ratios for each trait as a bar graph in order to confirm visually that at least a majority of the stereotypic traits have diagnostic ratios greater than one. This will confirm that our pre-survey method correctly selected for traits that are stereotypic for each social category. While this does not guarantee that our pre-survey method also correctly selected for non-stereotypes and counterstereotypes, it does make me feel better :) .

Tests A, B and C. I will then run three chi squared tests of independence between the diagnostic ratios of each of the traits: tests *A, B* and *C*.

Test *A* will compare the diagnostic ratios of the stereotypic traits and the non-stereotypic traits. This will simply serve as a baseline to ensure there is a difference between the diagnostic ratios of the stereotypic and non-stereotypic traits in my experiment, as I would expect. I should expect to reject the null hypothesis and accept the alternative hypothesis for this test:

HA_0 = There is no significant difference between the average of the diagnostic ratio of stereotypes and the average of the diagnostic ratio of non-stereotypes.

HA_a= There is a significant difference between the average of the diagnostic ratio of stereotypes and the average of the diagnostic ratio of non-stereotypes.

Test *B* will compare the diagnostic ratios of the stereotypic traits and the counterstereotypic traits. Again, I should expect to reject the null hypothesis and accept the alternate hypothesis.

HB_0 = There is no significant difference between the average of the diagnostic ratio of stereotypes and the average of the diagnostic ratio of counterstereotypes.

HB_a= There is a significant difference between the average of the diagnostic ratio of stereotypes and the average of the diagnostic ratio of counterstereotypes.

Test *C* will compare the diagnostic ratios of the stereotypic traits and the non-stereotypic traits. This is perhaps the most important test, as it will determine whether or not there is a significant difference between counterstereotypes and non-stereotypes .

HC_0 = There is no significant difference between the average of the diagnostic ratio of counterstereotypes and the average of the diagnostic ratio of non-stereotypes.

HC_a= There is a significant difference between the average of the diagnostic ratio of counterstereotypes and the average of the diagnostic ratio of non-stereotypes.

For all three tests, I will use $\alpha = 0.05$ in order to obtain 95 percent confidence. Since I am comparing two types of traits at a time and three traits for each type, I will use two degrees of freedom for each test.

Plotting the results. To go along with these chi-squared tests of independence, I should also plot the average diagnostic ratios of stereotypes, counter stereotypes and non-stereotypes along with some measure of variance/error. This will help to visually identify the presence of overlaps between the distributions of diagnostic ratios of stereotypes, counterstereotypes or non-stereotypes, and potentially characterize what I might estimate to be the typical “cut-off”

diagnostic ratios between all three categories of traits. An example would be to plot the averages as well as boxplots for each type of trait (Figure 6). While this figure is based purely hypothetical data, if these were the results I got, I would estimate that the cut-off between the diagnostic ratios of a stereotypic and counterstereotypic trait would be around 1.1, meaning that any trait that has a diagnostic ratio above this number is more likely to be a stereotype instead of a non-stereotype or counterstereotype. Similarly according to Figure 6, the cut-off between the diagnostic ratios of counterstereotypes and non-stereotypes seems to be about 0.75. This means that I would conclude that any trait with a diagnostic ratio below 0.75 is a counterstereotype, and traits with larger diagnostic ratios are most likely stereotypes or non-stereotypes.

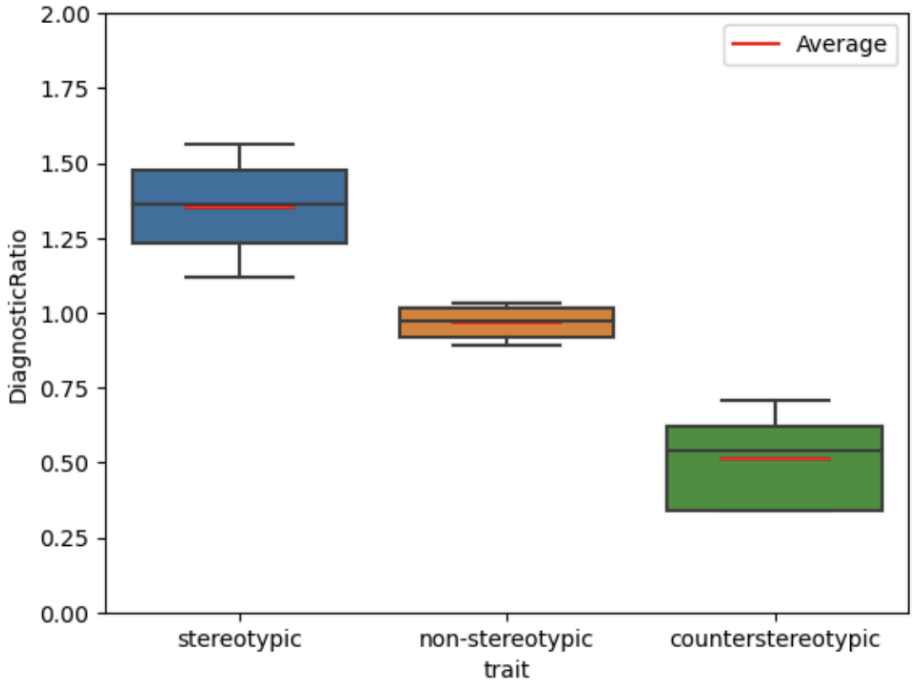

Figure 6: *Hypothetical plot comparing the averages and variance in diagnostic ratios of stereotypes, non-stereotypes and counterstereotypes.*

General Interpretations and Outliers. From the diagnostic ratios that we get from this experiment for each trait, we can make some general interpretations about the trait itself. While

highly unlikely, a trait with a diagnostic ratio of 1 would indicate that the population (of undergraduate students at MSU) believes that the trait is equally likely in both the social category and the general population. Therefore, the closer a trait's diagnostic ratio is to 1, the more likely it is to be perceived as being roughly likely in the general population and the social category.

It is also important to note the general trend that has already been characterized in the previous work on Bayesian probabilities. Traits with diagnostic ratios greater than 1 characterize the social category much more than the population, therefore they are stereotypes. The more a trait deviates from 1 in the positive direction, the 'stronger' of a stereotype the trait is for that social category.

However, we can also assume the opposite to be true as well. Traits with diagnostic ratios less than 1 (but greater than zero) characterize the social category much less than the population, therefore they are counterstereotypes. As the diagnostic ratio approaches 0, the 'stronger' the counterstereotype is.

Therefore, any outliers that we may find with diagnostic ratios that are very close to 0 are strong counter stereotypes, whereas outliers that are much greater than 1 are strong stereotypes.

Discussion

Why is characterizing counterstereotypes important?

The primary reason why it is important to figure the typical diagnostic ratios for counterstereotypic traits is to establish an empirical definition of counterstereotypes. While many counterstereotypes can be qualitatively described as being dictionary antonyms to stereotypic traits, this may be problematic for traits which do not have any clear antonyms, or those who have many loosely related antonyms. This study is meant to provide a framework for analyzing these more ambiguous traits.

Diagnosticity and Heider's Attribution Theory. Interpreting results of our experiment from the lens of an attribution perspective, knowing the diagnostic ratio of a trait may tell us how much a behavior may be perceived by the average person as being attributable to the social group that an individual belongs to. The more normative a trait or behavior is, the less individuating information that it provides about the individual. However, a trait or behavior that is not normative will provide a large amount of individuating information (Heider 1944). This idea may be applicable to the ideas of stereotypes, and counterstereotypes.

For example, an individual possessing a trait that is highly stereotypic for a social group that an individual belongs to may be seen as being normative, and so it will provide little individuating information about that individual in relation to their social group. A non-stereotypic trait could be considered normative for the general population (as well as the social group that a person belongs to), so little individuating information is gained from an individual possessing a non-stereotypic trait. However, an individual who possesses a counterstereotypic trait is not normative in relation to their social group, so a great amount of individuating information can be gained about the individual relative to their social group.

Knowing the attributions that people make for the traits that an individual possesses may be used to predict a person's behavior towards that individual. Numerous studies have shown that understanding the attributions that people make can be useful in understanding things such as their behavior or motivations (Seligman et al 1985, Lepper et al 1973).

Counterstereotypes and Perception. It has also been shown that individuals who possess counterstereotypic traits may be perceived in a more negative light compared to other individuals who belong to the same social category who do not possess these counterstereotypic traits, or even those who possess stereotypic traits (Ruben et al 2013). Therefore, identifying traits that are

seen as counterstereotypic may help individuals take proactive steps to prevent negative perceptions.

What are some limitations of this study?

I called this study a "pilot" study because there are some significant limitations to the findings that will need to be accounted for in later variations in order to make more generalized conclusions. For example, I will eventually need to test a greater number of social categories to ensure that my findings are consistent across more than just professional categories.

Additionally, taking inspiration from the Solanki and Cesario study, it may be useful to conduct follow-up studies on a more general population. This will allow me to extrapolate my results in differences of diagnostic ratios between counterstereotypes and non-stereotypes to general populations. As it stands right now, it would be uncertain if the same type of distinction I may observe between the diagnostic ratios of non-stereotypic and counterstereotypic traits in the undergraduate population also exists in the general population.

What are some additional things that may be interesting to test in the future?

Solanki and Cesario made the interesting point that people's motivation to control prejudice (MCP) may have played a role in their responses (Solanki and Cesario 2021). This may have potentially led to them collecting stereotypic traits that are not actually representative of the stereotypes that the undergraduate population holds about the social categories they tested. Although they framed their survey questions in a way to minimize this, I think it may have still played an effect, hence why there are some traits that were identified as stereotypes that still had diagnostic ratios that were less than 1. Therefore, in future iterations of this study (or perhaps as a second simultaneous study) I would like to change my pre-survey questions with the goals of coaxing out the "true" stereotypes of the undergraduate population, and to minimize the effect of

personal beliefs/stereotypes in the responses. In order to do this, I have devised a new methodology for the pre-survey section.

Modified Pre-Survey Protocol. In this new pre-survey, prior to the start of the questions in the consent form I will 'explain' that I have conducted an 'on-the-ground survey' in order to gather the options of a diverse sample of students on campus. The 'job' of the Ps is to play a family-feud style game so I can 'test' how good people are at predicting the opinions of others. If possible, I would also like to insinuate that Ps will get additional credit hours for their score from each 'correct' response. In reality, their score will be arbitrarily determined based off of the number of traits they provide, and all Ps will be granted the extra hours.

> **Modified Question:** We asked 100 (population) (insert question), and we put the top 10 answers on the board. In order to score the most points, you have to try to find the most popular answers to this question. You will not be penalized for wrong answers.
>
> **Example Modified Question using politician and Q1:** We asked 100 students on campus what they thought were common stereotypes for black people, and we put the top 10 answers on the board. In order to score the most points, you have to try to find the most popular answers to this question. You will not be penalized for wrong answers.

I deliberately have framed the questions in the style of the Family Feud game show ("..., and we put the top 10 answers on the board…find the most popular answers to this question…") in an attempt to influence the Ps to think about the stereotypes/counterstereotypes their peers may have. I said that Ps would not be penalized for incorrect responses to encourage Ps to list as many potential traits that they can think of. Traits that individual Ps provide that are not actually reflective of the options of the undergraduate population will inevitably be filtered-out since they are unlikely to have a large frequency when looking at the frequency of traits among all Ps.

This format of question can be applied to ask for both stereotypes and counter-stereotypes (Q1 and Q2). However in order to collect non-stereotypes, I would ask Ps a "bonus question" to boost their scores:

> **Non-stereotype question:** Congratulations, you've been selected for an opportunity to boost your score! As part of this bonus round, you will need to identify traits that you think are NOT considered to be counterstereotypic/stereotypic for (insert social category). An example would be the trait 'happy', as 'happy' is not considered a stereotype nor counterstereotype for (insert social category).

Similar to Q3 from the original pre-survey, I am unsure how useful this data will be, since the frequency in responses may not actually indicate the "best" non-stereotypes for a given social category.

What about 'weird' traits? Not all non-stereotypic traits are created equal. Personally, when I was coming up with traits that may be considered "non-stereotypic," I found myself not just thinking of traits that characterize both a social category and the general population, but also traits that characterize *neither* the social category nor the general population, abstract or unusual traits such as 'green' or 'spikey.' These traits are not typical predictions that we would make about other people in general, so I wonder if these traits will have diagnostic ratios close to 1 like other non-stereotypic traits, or if people's predictions will follow Bayes' Theorem at all.

On a similar note I think that it may also be interesting to traits that characterize everyone in the general population, such as 'breathing"' or 'alive,' perhaps as a control to ensure that Ps are answering the questions logically (the 'logical' probability response for all components of Bayes' Theorem would all be 100 percent, so I predict that all Ps who are answering the questions logically will have the diagnostic ratios for these traits be exactly 1).

Utilizing the unintended effects of counterstereotype interventions. Although counterstereotype interventions are largely ineffective (Moskowitz et al 1999, Galinski and Moskowitz 2007, Burns et al 2017), we can use some of the unintended consequences of them to our advantage. Since these interventions increase the accessibility of counterstereotypes and stereotypes for a social category, we could have all Ps complete a counterstereotype intervention prior to taking the pre-survey.

References

Introduction to Amazon Mechanical Turk - Amazon Mechanical Turk. (2017). Amazon.com. https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkGettingStartedGuide/SvcIntro.html

Jussim, Lee & Cain, T.R. & Crawford, Jarret & Harber, Kent & Cohen, Florette. (2009). The unbearable accuracy of stereotypes. *Handbook of prejudice, stereotyping, and discrimination*. 199-227.

Heider, F. (1944). Social perception and phenomenal causality. *Psychological Review*, 51(6), 358–374. https://doi.org/10.1037/h0055425

Karlins, M., Coffman, T. L., & Walters, G. On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology*, 1969,13, 1-16.

Katz, D., & Braly, K.W. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology*, 28, 280-290.

Kahneman, D., & Tversky, A. On the psychology of prediction. *Psychological Review*, 1973, 80, 237-251.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, 28(1), 129–137. https://doi.org/10.1037/h0035519

Lippman, W. (1922). *Public opinion.* New York, NY: Harcourt, Brace & Co.

McCauley, C., & Stitt, C. L. (1978). An individual and quantitative measure of stereotypes.

Journal of Personality and Social Psychology, 36(9), 929–940. https://doi.org/10.1037/0022-3514.36.9.929

Rubin, M., Paolini, S., & Crisp, R. J. (2013). Linguistic description moderates the evaluations of counterstereotypical people. *Social Psychology*, 44, 289-298. doi: 10.1027/1864-9335/a000114 Archived 2013-08-04 at archive.today

Seligman, C., Finegan, J. E., Hazlewood, J. D., & Wilkinson, M. (1985). Manipulating attributions for profit: A field test of the effects of attributions on behavior. *Social Cognition*, 3(3), 313–321. https://doi.org/10.1521/soco.1985.3.3.313

Solanki, P., & Cesario, J. (2021). Stereotypes as Bayesian Judgements of Social Groups. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43. Retrieved from https://escholarship.org/uc/item/23c8s4jp