Use Authorization

Signature ___________________________________

Date _____________________________________

Reliance on Stereotypes of White Non-Hispanic Cisgender Females and Males During Personality Judgment

Dissertation Proposal Document

by

Jacob Ralph Gibson

Idaho State University

A dissertation submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

in the Department of Psychology

Idaho State University

Summer 2022

To the Graduate Faculty:

The members of the committee appointed to examine the dissertation of JACOB RALPH GIBSON find it satisfactory and recommend that it be accepted.

Tera D. Letzring, Ph.D.
Major Advisor

Shannon Lynch, Ph.D.
Committee Member

Samuel Peer, Ph.D.
Committee Member

Jennifer McDonald, Ph.D.
Committee Member

Gesine Hearn, Ph.D.
Graduate Faculty Representative

Table of Contents

List of Figures

List of Abbreviations

Mturk	Amazon Mechanical Turk
BFI-2	Big Five Inventory 2
ISU	Idaho State University
J-Stereotype Consistency	Judge Stereotype Consistency
RAM	Realistic Accuracy Model
SAM	Social Accuracy Model
SRM	Social Relations Model
T-Stereotype Consistency	Target Stereotype Consistency

Reliance on Stereotypes of White Non-Hispanic Cisgender Females and Males During Personality Judgment

Abstract

There has historically been a plethora of research showing that personality judgments are generally accurate and that people do rely on stereotypes during general person perception. This is the first attempt to combine these areas of research and investigate how stereotypes impact the accuracy of personality judgments about specific persons. It was predicted that accuracy would be impacted both by how consistent an individual is with their group's stereotype and by how much the individual making the judgments relies on stereotype information. It was also predicted that individuals with less favorable views towards a group would use less individuating information and rely on their own negative stereotypes more, resulting in less accuracy overall. Three samples were recruited from both ISU and MTurk. One sample (*n* = 51) was used to create a stereotype profile for White non-Hispanic cisgender males and females based on items of the BFI-2. A second sample (*n* = 35) gave self-reports on their personality and were video recorded while talking about their life. These videos were then shown to the final sample (*n* = 209) who rated the personalities of those in the video. All ratings were compared against the stereotype profiles to assess stereotype consistency in personalities and judgments. Results supported the idea that accuracy is impacted both by how consistent an individual is with their group's stereotype and by how much the individual making the judgments relies on stereotype information, but also found that in general, judges who made judgments that were less consistent with stereotypes also made more accurate judgments. There was no evidence to suggest that favorability towards a specific group meaningfully impacted accuracy. Additionally, there was not a significant difference in how accurately judges rated male or female targets, or how stereotype-consistent male and female judges were. However, there was a significant difference how stereotype-consistent judgments were of male and female targets and a moderate effect in how stereotype-consistent male and female targets were. Overall, results indicate that those who, on average, made less stereotype-consistent judgments were more accurate.

Keywords: personality judgment, judgment accuracy, stereotypes, race, gender

Reliance on Stereotypes of White Non-Hispanic Cisgender Females and Males During Personality Judgment

When conflicts between groups arise, it is easy to point to stereotypes as the catalyst for between-group tension. People often believe that stereotypes are generally inaccurate and are overgeneralizations of a group (Jussim et al., 2015), and so it makes sense that stereotypes would take some blame for the conflict between different groups. A quick web search will show a plethora of dictionary definitions, social media posts, and research articles all decrying the inaccuracy of stereotypes (e.g., BetterHelp Editorial Team, 2022; Fiske & Durante, 2016). Unfortunately, it is rare for anyone to provide empirical evidence to back up the claims of stereotype inaccuracy (i.e., the idea that stereotypes are inaccurate overgeneralizations) that have been made (Jussim et al., 2015). In fact, research is increasingly showing that, at the group level, stereotypes have medium to high levels of accuracy, and stereotype inaccuracy is the exception (Jussim et al., 2015, 2016; Mackie, 1973; Ryan, 2003). The prevalence of stereotype accuracy means that using stereotypes as a heuristic to make judgments about others should help individuals to make more accurate judgments when there is a lack of individuating information about the person being judged. Some researchers argue that even if stereotypes are accurate, using them to make judgments is unfair to the individual being judged because stereotypes do not represent every member of a group (Fiske, 1989; Stangor, 1995, 2016). This idea has some merit because although on the group level stereotypes may be accurate, a significant amount of research shows greater variability within groups than there is between groups (See Ellemers, 2018 and Hyde, 2014 for a review). This means that even accurate stereotypes will not fully represent every member of a group (and because they are an average, they may not fully represent any member of the group), but they should generally represent at least some aspects of most members of a group and lead to greater accuracy compared to using no information.

Thankfully, stereotypes are not the only pieces of information that people use when making judgments of others (e.g., people also use physical appearance and various behavioral cues; Funder, 1987; Latif et al., 2022; Naumann et al., 2009), and research has shown that in general, people are usually fairly accurate in many of the personality judgments that they make (Letzring & Funder, 2018). The aim of this research project is to increase understanding of the extent to which people use stereotype information when making personality judgments of others (specifically how stereotypes impact the accuracy of personality judgments), and under what conditions people ignore any individuating (unique, distinctive) information about someone when making these judgments.

Introduction

Even from a very young age, humans begin to form associations that help them to identify and categorize different aspects of the world around them (Packer & Cole, 2015; Quinn et al., 2002; Waxman & Gelman, 2009). Through this process, individuals soon learn to regard objects that look, feel, or act similarly as being in a similar category, and this information is used to help us understand and work within the world. The paper or screen you are reading this on is, in many ways, similar to other papers or screens you have encountered. Most likely, it did not take a lot of work for you to recognize the object you are using because you have had experience with similar objects in the past. Even with this experience, there are likely some differences between the current object you are looking at and those you have seen before. Our world is inherently variable, and even two things that look and act in almost the exact same way will likely have some differences. This means that using category membership in order to understand an object will help us gain a guess of what to expect from a given category. It also means that individuals have to be flexible in their understanding and realize that, although there may be many prototypical examples of a category (a chair with four legs and a back), there may be many instances where something does not fit perfectly into a category (e.g., a bean bag chair may still be considered a chair; Smith & Zarate, 1992).

When making judgments of other individuals' personalities, many of the same processes that individuals use when categorizing and judging objects may be at play. All individuals form categories of various groups and some of this information is about how members of a specific group act on average (i.e., what the personalities of group members are generally like). If this information is accurate, then relying on these categories when making judgments of others should help individuals have a basic idea of what to expect from a given group member. On the other side of this, if inaccurate or false information about a group is used, it will lead to false perceptions and an incorrect understanding of what to expect from a given group member.

Traditionally, these social categories about groups of people have been called stereotypes. Researchers have defined stereotypes in many ways over the years. Most have defined them as inaccurate (Jussim et al., 2015), and say that they help to rationalize prejudice (LaPiere, 1936), that they are an inaccurate reflection of reality (Bargh & Chartrand, 1999), or that they represent small kernels of truth which have been exaggerated (Allport et al., 1954). The American Psychological Association (1991), in a United States Supreme Court case involving potential discrimination, defined stereotypes as "overgeneralizations [that] are either inaccurate or do not apply to the individual group member…" (p. 1064). All of these definitions would be perfectly acceptable if the term "stereotype" was exclusively used with reference to overgeneralized and/or inaccurate beliefs about a group of people. Unfortunately, decades of researchers have freely used the term stereotype to refer to any belief about any group and have almost always failed to provide any evidence supporting claims of overgeneralization or inaccuracy (see Jussim [2012] for a comprehensive review; see Jussim et al. [2018] for an explanation of why many researchers have ignored the idea of stereotype accuracy). If all beliefs about all groups are stereotypes, and all stereotypes are inaccurate, it means that any beliefs about any group would be inaccurate, and it would be impossible to have accurate beliefs about any group (Jussim et al., 2015).

For these reasons, if researchers define stereotypes as inaccurate overgeneralizations about a group, they need to limit themselves to only those group descriptions that have been demonstrably inaccurate or overgeneralized (this, unfortunately, is rarely done; Jussim, 2012). When the ideas of overgeneralization and inaccuracy are dropped, and stereotypes are instead defined simply as people’s beliefs about groups and their individual members (Ashmore & Del Boca, 1981), research has found that inaccurate stereotypes are the exception and that most stereotypes have moderate to high levels of accuracy (Campbell, 1967; Jussim et al., 2016, 2018, 2021; Mackie, 1973; Ryan, 2003). This means that (like with any other form of category) if stereotypes are accurate, they should represent a generalized belief that is at least partially accurate for most members of a group most of the time. This also means that, when there is a dearth of individuating information (information about the unique aspects of another person), relying on a stereotype to make assumptions about a group member should lead to more accurate predictions more often than if stereotypes are not used. It also means that if a stereotype is inaccurate, it should not be relied on because it will lead to inaccurate assumptions most of the time. Although this makes practical sense, there is currently no research demonstrating this effect for judgments of specific individuals.

This is an important avenue for research because stereotypes–when defined as beliefs about groups and their individual members (Ashmore & Del Boca, 1981) or as general expectations about members of a particular group (Ellemers, 2018)–are an integral part of social interaction. The tendency, shown earlier, to assume stereotypes are inaccurate has likely had the unfortunate consequence of disincentivizing research investigating the potential utility of accurate stereotypes (and the potential harm). Research has continually emphasized the importance of first impressions (Anderson, 1965; Anderson & Barrios, 1961; Asch, 1946; Berkowitz, 1960; Latif et al., 2022; Lorenzo et al., 2010). Research has also demonstrated that the initial impressions that are formed may impact the accuracy of future judgments (Gibson, 2019), and this has consequences for relationship development (Human et al., 2013). When individuals come to an interaction with no other information about someone's personality, they may rely on stereotypes to help them make their initial judgments, but little is known about what impact stereotypes have on the accuracy of personality judgments of individuals.

Recent research has identified a few questions that, up until recently, have largely gone unasked in relation to stereotypes and stereotype accuracy (Jussim et al., 2018). Two of these questions are particularly relevant to this project. They are:

1. "When and how does relying on a stereotype increase the accuracy of person perception?" (Jussim et al., 2018, p. 214) and

2. "Do people ever actually ignore individuals' personal characteristics when perceiving, evaluating, and judging them?" (Jussim et al., 2018, p. 214)

The purpose of the current project was to begin to answer these questions within the framework of personality judgment accuracy. This is an important area of study because much of our daily interactions involve making judgments of others, but there is not a lot that is known about the extent to which stereotypes impact this process. A significant amount of research on this topic has been done within social cognition, but to my knowledge, no one has investigated the impact that stereotypes have on the accuracy of personality judgments of individuals.

Social Cognition, Bias, and Accuracy

Much of the social cognition literature is replete with the assumption that stereotypes lead to biased thinking and self-fulfilling prophecy in social interactions (Jussim, 2012; Operario & Fiske, 2004). This is because, as was said earlier, stereotypes have often been assumed to be inaccurate reflections of a group (Jussim et al., 2015) even where there is no evidence for the inaccuracy of the stereotypes mentioned. The basis behind many of these assumptions comes from the idea that the processes that lead to biased judgments and errors in the lab will also lead to biased judgments and errors in regular social interactions (Jussim et al., 2005). The problem with this view is that most of the time when the person perception process is tested in a lab setting, a participant is asked to give a rating of a fictitious person. In these cases, it is usually considered an error in judgment if the participant uses stereotypes, or if they are susceptible to primed information when making a judgment. These lab scenarios may lead to errors in a lab situation, but some researchers have argued that these same processes may often lead to more accurate judgments in regular everyday social interactions (Blackman & Funder, 1998; Funder, 1987; Jussim, 2012).

Indeed, there is now a large body of research showing that, in general, people can be reasonably accurate in their social judgments despite the many errors they have been shown to make in lab situations (for a review, see Letzring & Funder [2018]). Although there has often been a division between the research of social cognition and that of personality judgment accuracy (Uleman & Kressel, 2013), it is possible to reconcile these differences, and researchers have attempted to do just this to varying degrees (Jussim et al., 2016; Zaki & Ochsner, 2011). In order to understand why there are differing conclusions and how they may be reconciled, it is vital first to understand the history behind the field of personality judgment accuracy and social cognition.

History of Personality Judgment Accuracy

Accuracy research can be traced back to individuals such as Adams (1927), Allport (1937), and Taft (1955), who investigated the attributes that would make someone a good judge (someone who is consistently accurate in their impressions of others). In general, they found that some aspects of personality, such as intelligence and social skills, were related to more accurate judgments on average. Although there was promise in this area of research, a critique by Cronbach hindered research on accuracy for a time. Cronbach's (1955) critique was that research on accuracy was ignoring some critical methodological concerns that needed to be addressed. In this critique, he claimed that accuracy scores reflected several different components that, if studied using a single score, were confounded and made accuracy scores difficult or impossible to interpret. For example, research at this time demonstrated that people often judge others as similar to themselves (known as *assumed similarity*), and so when accuracy researchers studied personality judgments as a single score, it confounded the result by combining judgments of assumed similarity with any distinctive (unique) judgments (Uleman & Kressel, 2013). This made it impossible to tell if people were actually accurate at judging others, or if this was just an artifact of judging others as being similar to themselves.

Cronbach proposed that accuracy scores should be decomposed into several different components so that the different interacting forces on accuracy could be studied independently. This would make it possible, for example, to determine the extent to which a judge (the person making the personality judgments) was indeed accurate at judging a target's (the person for whom personality judgments are being made) unique aspects, or if the judge was simply rating a target in a way that was favorable. For example, Cronbach suggested one component of accuracy called *elevation accuracy*, which measures a judge's tendency to over or underestimate all targets that they encounter (when using self-report measures, this score would actually reflect response bias or a judge's tendency to default to a specific rating on a scale; Cronbach, 1955; Jussim, 2012; Kenny, 1994)

Cronbach also suggested a component called stereotype accuracy (this is different from the stereotype accuracy mentioned earlier), which essentially is a measurement of trait effects among a set of targets. Cronbach's stereotype accuracy compared a judge's rank order of traits among a set of targets to the observed average rank order of traits among a set of targets. The closer a judge's rating is to the observed average rating, the higher their stereotype accuracy. For example, a judge may consistently rate the trait of extraversion as being more prevalent among a set of targets than the trait of conscientiousness. Although Cronbach used the term stereotype accuracy, this is unrelated to what most people think about when they think of stereotypes such as race, sex, gender, or religion. It also has little to do with the current study of stereotype accuracy discussed previously or with how stereotypes are defined in this paper. Cronbach also suggested differential elevation accuracy (how accurately a judge can rank order targets on each trait, or a measure of target effects) and differential accuracy (after controlling for all other components of accuracy, how accurately a judge can rank order a set of targets averaging across all traits).

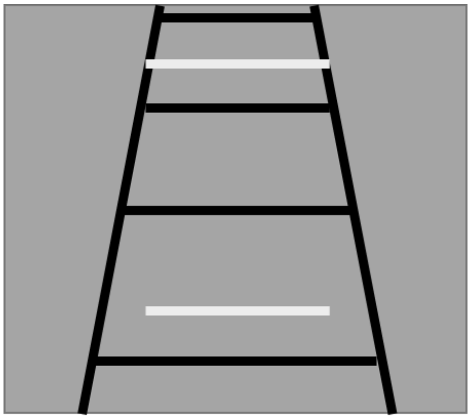
Cronbach's critique showed that studying accuracy as a single score may be misleading, and so future researchers were encouraged to study accuracy in a way that accounts for the various components Cronbach outlined. It is possible to compute Cronbach's components of accuracy using a two-way analysis of variance with target and trait as factors, but at the time of the critique, it may have seemed too difficult or tedious for most researchers to address (Funder, 1987; Jussim, 2012; Kenny, 1994). The result of this critique was that the study of accuracy receded while the study of general person perception and social cognition became more popular. These areas bypassed the accuracy questions by investigating the process (instead of the results) of making errors and being biased in social judgments (Funder, 1987; Uleman & Kressel, 2013; Zaki & Ochsner, 2011). These areas typically use hypothetical individuals instead of actual people, and so the question of accuracy is not addressed. Doing this made it possible to study the process of person perception without having to worry about the difficulties identified by Cronbach.

RAM and The Return of Accuracy Research

The decades after Cronbach's critique led to a significant amount of important research on the social and cognitive processes underlying person perception. This research has helped to better understand the "how" of interpersonal perception. The downside is that this research was conducted with a potential cost to external validity. Without investigating the accuracy questions, it is difficult to know if the findings that resulted from research using hypothetical individuals within a lab would extrapolate to real-world interactions. To help elucidate this point, consider the classic linear perspective illusion (see Figure 1). When looking at a two-dimensional image of a train track that starts near the bottom of an image and then converges in the distance as the tracks move vertically, it appears that the railroad ties that are closer to the bottom of the image are the same size as the ones closer to the top. This, of course, is an illusion caused by representing a three-dimensional image in only two dimensions. Although this leads to an error

Figure 1

Classic Linear Perspective Illusion



in a lab situation, it often leads to an accurate perception in most real-life situations (i.e., most railroad ties are relatively the same size along a track). This analogy helps clarify how processes that lead to reliable perceptual errors in a lab setting may actually lead to accurate judgments in real-world situations.

For this and similar reasons, some individuals began calling for research on accuracy to return (Funder, 1987, 1995; Kenny & Albright, 1987). A few models helped bring about the return of accuracy research. One such model was the social relations model (SRM; Kenny, 1994; Kenny & La Voie, 1984), which can be used to study accuracy using components that were similar to those advocated by Cronbach (Uleman & Kressel, 2013). One difference between the SRM and Cronbach's components was that Cronbach's components were designed to investigate the ratings of a single judge on multiple traits and multiple targets. Cronbach's analysis would be rerun for each judge in a study. The SRM, on the other hand, was designed to investigate the ratings of multiple judges and multiple targets on a single trait. SRM analysis would be rerun for each trait under investigation. This helped to ignite a considerable amount of research because it provided a framework for assessing questions about social perception that often were not previously addressed (Kenny et al., 2006).

The realistic accuracy model. Another model that significantly helped the return of accuracy research was one that conceptualized the accuracy process in a way that was relatively straightforward and easy to follow and test. This is known as the realistic accuracy model (RAM; Funder, 1995). This model is based on the Brunswik lens model (a model designed to explain the process of perception through the use of object cue relevancy and utilization; Brunswik, 1956). The RAM articulates four steps that must be completed in order for a judge to reach an accurate perception of an aspect of a target. The four steps described by the RAM are separated into two steps that are specifically about a target's behavioral cues (i.e., behavioral actions that can be associated with an aspect of a target's personality) and two steps about a judge's perceptual process. According to the RAM, in order for a judge to reach an accurate judgment of an aspect of a target, a target must first exhibit cues that are *relevant* to the trait being judged, and these cues need to be *available* for the judge to *detect* and *utilize* in making a judgment.

To help clarify this process, imagine the following example about an agreeable target. In order for a judge to accurately perceive a target's agreeableness, the target must first have behavioral cues that are *relevant* to the trait of agreeableness, such as being polite and willing to work with others. These cues must then be outwardly expressed and *available* as behavioral actions. For example, a target may react positively to a judge and may offer to help them or work with them. A judge then needs to *detect* these agreeableness-related cues by being in a situation where they can pick up on the cues and by paying attention to them. Finally, the judge needs to correctly *use* these behavioral cues by attributing them to a target's underlying disposition to be agreeable instead of misattributing them to various situational forces or to a target's desire to only appear agreeable (such as during a job interview or when being evaluated). According to the RAM, for an accurate perception to be reached, all four steps need to be completed entirely and in order. If this process is frustrated at any point, it will be unlikely for an accurate perception to be reached by a judge.

Along with breaking down the accuracy process into steps that could be understood and tested, early research on accuracy identified four moderators that, within the framework of RAM, help to articulate factors that influence levels of accuracy. One large advantage of these moderators is that they have the ability to explain both accurate and inaccurate judgments (Letzring & Funder, 2021). The first of these moderators have already been addressed and is known as the *good judge*. Some research has suggested that there can be considerable variability across judges (Colman et al., 2017), which indicates that it might be important to understand what makes someone a better or worse judge of personality. It was a desire to find the characteristics of a good judge that drove much of the early work on accuracy (e.g., Adams, 1927). Research on the good judge has identified a number of characteristics that are correlated with accurate personality judgments. One characteristic of a good judge is their level of cognitive functioning. Research has demonstrated that aspects of cognitive functioning such as intelligence (Davis & Kraus, 1997; Harris et al., 1998) and attention (Biesanz et al., 2001; Waggoner et al., 2009) may account for some of the variability between judges. In addition to cognitive functioning, aspects of a judge’s personality may also impact the accuracy of their judgments. Research that conceptualizes personality using the Big Five traits (John et al., 2008) has demonstrated that a judge's level of agreeableness, conscientiousness, and emotional stability may moderate their levels of accuracy (Hall et al., 2016; Letzring, 2008). Additionally, a judge's empathetic tendencies may enable them to more fully understand the relation between a target and their behavior and therefore make more accurate personality judgments (Colman et al., 2018; Hall et al., 2016). Finally, aspects such as motivation (Biesanz & Human, 2010) and social skills (Letzring, 2008) may help judges to elicit more relevant cues from targets. There is still a lot of work that needs to be done to help elucidate what makes someone a good judge of personality, but it seems clear that some people are generally better equipped to make accurate judgments (for a comprehensive review of research on the good judge, see Colman [2021])

The second moderator of the RAM is known as the *good target*. Just as some judges are more accurate on average when compared to others, some targets are more accurately judged on average when compared to others. Several different characteristics have been shown to be related to being a good target of personality judgments. First is a target's level of psychological adjustment. Targets who are more psychologically adjusted (e.g., higher well-being and self-esteem, and lower depression) tend to be rated with greater expressive accuracy (a measure of how accurately, on average, a given target is rated by a set of judges; Human et al., 2019). This is possibly because these individuals tend to exhibit cues that are more relevant to their personality (Human et al., 2014). In addition to psychological adjustment, good targets also tend to behave in predictable ways based on the behaviors expected by their personality. This is known as personality coherence (Cervone & Shoda, 1999), and research has demonstrated that it may be related to a target being rated more accurately (Biesanz & West, 2000; Human et al., 2014). Some research also suggests that part of what may make someone a good target is having good self-knowledge. These individuals have strong self-concept clarity (i.e., a clear idea of who they are and where they want to go in life; Lewandowski Jr et al., 2010) and they may be reflective about their motives, thoughts, and feelings (Mignault & Human, 2021; Scheier et al., 1978). Finally, good targets tend to be those who have good social skills or who are able to make good impressions (Human et al., 2012), and those who are physically attractive (Lorenzo et al.,2010; see Mignault and Human [2021] for a comprehensive review on the good target).

The RAM also posits a third moderator known as the *good trait*. Like judges and targets, some traits are typically more accurately judged when compared to others. A number of factors have been shown to impact how accurately a trait is judged. The first of these is the observability of a trait. Based on the RAM, traits that exhibit more relevant and available cues should be judged with greater accuracy. A common example of this is the trait of extraversion, which is typically one of the most observable traits because it is outwardly expressed. Compare this to neuroticism, which is a measure of the prevalence of negative emotions and is not typically a highly expressive trait. Overall the observability or visibility of a trait is one of the most well-supported moderators of how accurately a trait can be judged (Kenny & West, 2010). Favorability or social desirability is another factor that researchers have suggested may impact the accuracy with which a trait can be judged. Research on the impact of favorability on accuracy has resulted in mixed findings with some research finding the two to be positively related (e.g., Funder, 1980; Funder & Colvin, 1988) while others have not found a relation (e.g., Paunonen & Kam, 2014; Ready et al., 2000). There are probably many factors at play that impact the relation between favorability and accuracy, and more research needs to be done in this area (Krzyzaniak & Letzring, 2021). Along with specific trait factors, external forces have also been shown to impact trait accuracy. For example, in anxiety-provoking situations, neuroticism can be judged with greater accuracy, likely because there are more cues available to the judge (Hirschmeüller et al., 2015). Finally, the perspective of the judge can impact which traits are salient and therefore judged with greater accuracy. Research on the self-other knowledge asymmetry model (Vazire, 2010; see Bollich-Ziegler [2021] for a review of this model) has demonstrated that some traits are judged better by the self, while some traits are judged better by others. This is because the self and others each have different cues readily available to them and self-judgments tend to skew favorably on highly evaluative traits. This means that the self and others will see different aspects of traits and focus on different aspects of personality that help them to be better or worse at judging that trait (for a comprehensive review of research on good traits see Krzyzaniak and Letzring [2021]).

The final moderator proposed by the RAM is *good information*. Good information is typically broken down into two parts, quantity and quality. The quantity aspect of good information refers to the fact that the more behavioral cues a target exhibits for a judge to use, the more accurately a target is usually perceived. In support of this, research has found that generally, as the length of acquaintanceship between a judge and target increases, so does accuracy, even when assumed similarity is accounted for (Brown & Bernieri, 2017; Funder et al., 1995). Although accuracy may increase with more information, not all information is created equal, and so the quality of the additional cues plays an important role in the accuracy of judgments. This is because cues that are more indicative of a specific trait, or more directly relevant to a trait, should lead to greater accuracy. An example of this can be seen in the fact that targets are judged with greater distinctive accuracy (i.e., judgments of the unique aspects of an individual) when a judge hears them talk about their thoughts, feelings, or behaviors compared to when they are seen engaging in various behaviors (e.g., reading a poem, telling a story; Letzring & Human, 2014). This is because a direct discussion of someone’s thoughts and feelings provides more directly relevant information than simply observing someone, and provides judges with information they otherwise cannot access (see Beer [2021] for a review of research on the good information moderator of RAM).

All of the moderators of RAM have been shown to play an important part in the accuracy process. The RAM as a whole has been a vital addition to research on personality judgment accuracy. Recent years have seen the addition of a new computational model that has helped to change the way accuracy research is conducted and analyzed.

The Social Accuracy Model

Although the SRM addressed many of the criticisms of Cronbach, some features were still not fully addressed. SRM worked mostly on a trait level, could only investigate a single trait at a time, and focused more on consensus than accuracy. This made it difficult to test accuracy simultaneously across traits. A more recent computational model that has gained increasing use is the social accuracy model (SAM; Biesanz, 2010). The SAM decomposes accuracy scores into judge effects known as perceptive accuracy (i.e., how accurately, on average, a judge rates a group of targets), target effects known as expressive accuracy (i.e., how accurately, on average, a target is rated by a set of judges), and dyad effects known as impressionistic accuracy (i.e., how accurately a single judge is of a single target). Each of these effects can be further divided into normativity (also known as normative accuracy) and distinctive accuracy. Normativity addresses the extent to which judges rate targets (or the extent to which targets are rated by judges) to be like the average or normative person. Normativity is typically calculated by comparing judge ratings of targets against a normative profile that is often created by averaging together all ratings of all targets (this is sometimes divided by binary gender, which was done in the current study) to get an idea of what the average person is like. The result of this process is a score for each judge that shows how normatively a judge rates a set of targets (i.e., perceptive normativity) and a score for each target that shows how normatively a target is rated by a set of judges (i.e., expressive normativity). Distinctive accuracy is estimated by subtracting the normative profile from each target's personality profile in order to create a distinctive personality criterion for each target that captures the ways that each target deviates from the normative profile. Just as with the ratings on normativity, judge ratings can then be compared against each target's individual distinctive profile in order to ascertain how accurately on average a judge rates the distinctive aspects of targets (i.e., perceptive distinctive accuracy) or on average how accurately the distinctive aspects of a target are rated by a set of judges (i.e., expressive distinctive accuracy).

A plethora of accuracy research has recently used the SAM because it has many advantages over previously used computational approaches to accuracy. For example, because the SAM is a multilevel model, studies that are designed with judges nested within targets, and targets nested within judges (which better approximates how normal daily interactions work) can obtain more reliable estimates of accuracy than was previously possible. The SAM can account for this non-independence, thus allowing for a more pragmatic testing of the social perception process. SAM also makes it relatively easy to test various moderators of the accuracy process such as the impact of prior true or false information on normativity and distinctive accuracy.

The impact of previous information and confirmation bias on personality judgments

One phenomenon that complicates this research is that often people fall prey to confirmation bias (Oswald & Grosjean, 2004) where they selectively attend to information that confirms their current beliefs while ignoring or devaluing information that is inconsistent with their beliefs. This seems to be more likely in situations where individuals hold their beliefs with greater conviction (Hart et al., 2009). Research has also shown that negative personality-relevant information (information about undesirable personality traits) often is processed more thoroughly than positive information (Baumeister et al., 2001; Pratto & John, 1991) and that negative information is more resistant to retroactive interference (changing previously learned information in favor of new information) compared to positive information (Ybarra, 2001). Research also suggests that negative information about others is more readily attributed to dispositions while positive information is often attributed to situations (Reeder & Spores, 1983; Rothbart & Park, 1986; Ybarra et al., 1999). Taken together, these results suggest that negative evaluations of a group will be processed more thoroughly, be more readily attributed to a group's dispositions, and be held with greater conviction. The result of this is a greater confirmation bias among individuals who have a negative stereotype of a group. Understanding what impact this has on the accuracy of those judgments is still an open question and one of the purposes of this project.

Stereotype Accuracy and Personality Judgment

Previous research has delineated three steps that must be taken in order to validate the accuracy of stereotypes (Jussim et. al., 2019, 2015). First, researchers need to investigate what beliefs people have about a specific group of interest (e.g., do men tend to be less open to experience than women?). Next, researchers need to find or create a criterion against which this belief can be compared (e.g., find or create a large representative dataset that assesses the personality traits of men and women). Finally, researchers need to compare the beliefs people hold about this criterion. If there is a high degree of correspondence between beliefs and criterion, then researchers can confidently claim they have demonstrated the accuracy of the investigated stereotype.

When using this method, research has demonstrated that stereotype accuracy is a very large and replicable effect (possibly one of the largest in all of social psychology; Jussim et al., 2016). This does not mean that every member of a group perfectly fits a stereotype. Instead, it means that, if a stereotype is accurate, it represents a group as a whole and the mean attributes of people within that group. This means that if a judge has no other information about a given target, relying on judgments consistent with an accurate stereotype should lead them to reach greater accuracy more often than when not using a stereotype or if they make judgments that are inconsistent with the stereotype. This also means that when a judge has no other information about a given target, relying on judgments consistent with an inaccurate stereotype, should (based on the research cited earlier about the negative impact of false information on levels of accuracy) lead them to reach lower levels of accuracy compared to not using a stereotype or if they make judgments that are inconsistent with the stereotype.

Evidence is slowly starting to accumulate to support the accuracy of many stereotypes (Crawford et al., 2011; Jussim et al., 2005, 2015, 2016). Most research on the accuracy of personality stereotypes of race, gender, regional character (e.g., western or eastern United States) and national character (e.g., Germans or Canadians) use data from various studies that assess personality using the Big Five inventory or the NEO-PI (or newer variations). These inventories have been given to thousands of individuals around the world and often act as a criterion against which personality perceptions can be compared. In a recent review of the accuracy of stereotypes about personality (Jussim et al., 2021), stereotypes about the personalities of different nationalities and regions were generally low in accuracy, while those about gender and age were high in accuracy. The authors suggest that this might be because people generally have a lot of interactions with people of varying genders and ages, but do not have a lot of interactions with people from varying nationalities. Although this review collected some interesting findings, there is still a significant amount of work that needs to be done in order to understand the accuracy of stereotypes about personality. This work will be vital to helping to answer the two important questions mentioned earlier about person perception, namely when and how relying on a stereotype may lead to an increase in the accuracy of personality judgments and under what conditions people ignore an individual’s personal characteristics when perceiving, evaluating, and judging them (Jussim et al., 2018).

Use of individuating information

There is some work within person perception research that may suggest important avenues of investigation when trying to determine the conditions under which people ignore individuating information. Research has shown that there is a connection between psychological essentialism (i.e., believing that there is an underlying essence or innate aspect of a category that impacts the members of that category; Gelman, 2004) and racial prejudice (Mandalaywala et al., 2018). This latter research suggests that as essentialist thinking increases, there is greater endorsement of existing social structures and hierarchies among both Black and White participants. Although prejudice and essentialist thinking are related, it is important to note that essentialist thinking itself may not be the cause of prejudice. Research has found that essentialist thinking does not always lead to increases in prejudiced thinking and de-essentialist thinking does not always lead to decreases in prejudiced thinking (Haslam et al., 2002; Verkuyten, 2003). This is consistent with the arguments made so far in this paper because if stereotypes are generally accurate (which the research presented so far supports) there must be some underlying essence that unites members of a group such as genetic and/or cultural factors. The accuracy of stereotypes may come from a general ability to correctly identify and use information about the uniting essence of a group.

It may be that essentializing a group is not the issue but instead, it is the negative connotation associated with the perceived essence of a group. As stated earlier, negative information is more resistant to change. Research has also shown that negative information tends to be more salient compared to positive information (Fiske, 1980; Pratto & John, 1991; Vaish et al., 2008). Individuals who consistently devalue members of a specific group may be those who, for various reasons, are unable or unwilling to work to inhibit their negative views of a group and therefore are more likely to prioritize this information when making judgments of others. This may also cause them to be more rigid in their essentialist views of a group and may degrade their ability or desire to use individuating information when making judgments of group members. To date, this is not something that has been investigated within the field of personality judgment accuracy.

Accuracy of Gender Stereotypes

When stereotypes are defined as people's beliefs about groups and their individual members (Ashmore & Del Boca, 1981), both accurate and inaccurate stereotypes of groups are encompassed in a single idea. Individuals may create a stereotype for any sized group, making it possible to create an almost limitless number of stereotyped groups. It is impossible to test all the potential stereotype variations within a single project. For this reason, the current project was limited to stereotypes that have been shown to be generally accurate, specifically stereotypes about gender.

A recent annual review of psychology paper suggested that although gender stereotypes may hold a kernel of truth (the idea that there may be small amounts of truth to them) this does not outweigh the overabundance of inaccuracy contained in them (Ellemers, 2018; see Jussim [2018] for a critical review). However, this review almost never cites any research that directly tested accuracy, and it fails to mention a few other errors. For example, one of the papers it does cite (i.e., van der Lee & Ellemers, 2015) claims gender was predictive of the likelihood of getting funding for research (with women being less likely), but this is better explained by the fact that women also applied for funding from sources that are less likely to award that funding (Albers, 2015), but only the original claim was mentioned by the review paper. Finally, the review paper did not mention or address the plethora of research supporting the accuracy of gender stereotypes. For example, research has found that people have gender stereotypes that lead to medium or high levels of accuracy about academic GPA (Beyer, 1999), professional pursuits and advancement (Cejka & Eagly, 1999; McCauley et al., 1988, 1981; McCauley & Thangavelu, 1991), performance on various cognitive tasks (Halpern et al., 2011), nonverbal communication (Briton & Hall, 1995), and a host of personality traits and behaviors (Allen, 1995; Hall & Carter, 1999; Löckenhoff et al., 2014; Martin, 1987; Swim, 1994). This last characteristic, personality traits, is highly important for the current project. These results all support the idea that there is a significant amount of utility in using stereotypes when making judgments.

The Current Project

Research has shown that directly giving judges true information about a target's personality does not seem to make a significant difference in distinctive accuracy (compared to having no additional information; Gibson, 2019) but does lead to small decreases in normativity. One reason for this finding is that the personalities of many individuals align with the groups to which they belong. The groups people are born into likely impact their personalities from a young age, and their personalities help form what groups they chose to join later in life. This may be one factor that leads to the accuracy of stereotypes (Jussim et al., 2016). This means that judges already have relevant information about a target (in the form of previously learned stereotypes), and using that information should lead to a significant increase in distinctive accuracy. A main point of this research is to try to understand if this is indeed true. Do stereotypes (specifically those about gender and personality) have a meaningful impact on personality judgments?

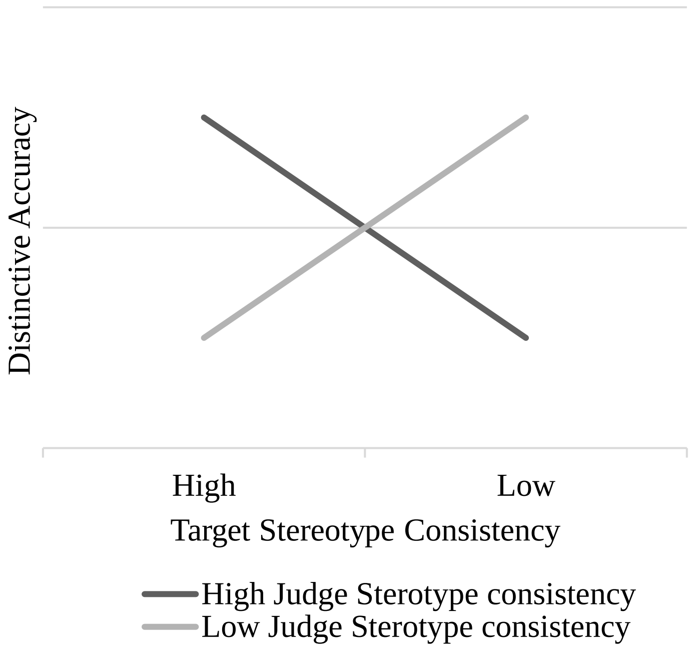
In order to answer this question (and those mentioned earlier), this study will use the SAM as an analytical model. No research to date has used SAM to investigate the impact of stereotypes on the personality judgment process or the relations between stereotypes and normativity or distinctive accuracy. As mentioned, normativity is typically calculated by comparing a judge's rating of a target to a normative profile. This profile is created by averaging together the personality profiles of a number of individuals (sometimes further divided by binary gender, which was done in the current study), which is conceptually similar to the process of assessing stereotype accuracy. Within stereotype accuracy, a criterion is created that represents the mean levels of a group, and then the beliefs of a group of judges are compared to this stereotype criterion in order to ascertain the accuracy of their stereotypes. The current project uses the term *stereotype consistency* to refer to how correlated a target's individual personality profile is with the stereotype profile for their gender.

With this in mind, this project has two main hypotheses designed to address the two unasked questions mentioned at the beginning of the introduction section:

1. A judge's level of distinctive accuracy will be moderated by a target's level of stereotype consistency and by a judge's utilization of stereotype information (see Figure 2).
    a. Judges who make more stereotype consistent judgments will be more distinctively accurate when targets' personalities are more consistent with their stereotype profile than when targets' personalities are less consistent with their stereotype profile. On the other hand, judges who make less stereotype consistent judgments will be more distinctively accurate when targets' personalities are less consistent with their stereotype profile than when targets' personalities are more consistent with their stereotype profile.

Figure 2

Representation of Expected Results for Hypothesis 1

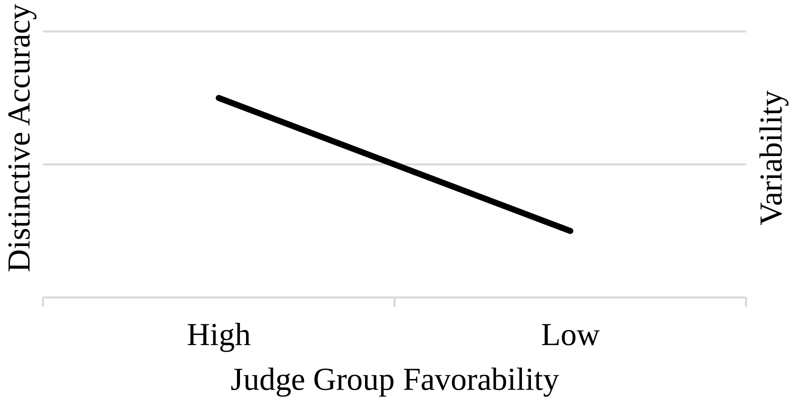


2. Individuals with greater favorability, or those who have a more positive perception of a specific group, will be less likely to ignore individuating information about specific targets and therefore will judge targets from that group with greater variability and higher levels of distinctive accuracy (see Figure 3).

Hypothesis 2 is based on the assumption that judges who have more negative evaluations of a group will have greater confirmation bias, causing them to attend more to those cues that are in line with their negative stereotype while ignoring or devaluing cues that are inconsistent with their views of the target group. This means that a judge's positive or negative evaluation of a group will moderate their level of distinctive accuracy because judges with negative evaluations will use less individuating information, which will lead to lower distinctive accuracy. This also means there will be less variability in how judges with less favorable judgments will rate targets (compared to more neutral or favorable judgments) because judges with less favorable judgments will rely more on their personal stereotypes and less on what the target is actually like.

Figure 3

Representation of Expected Results for Hypothesis 2



In addition to these hypotheses, the current project also investigated a few research questions that do not have any specific predictions. Finding answers to these questions may go a long way towards increasing understanding of how stereotypes are used in the personality judgment process.

1. Are there any differences in normativity or distinctive accuracy for male versus female targets?
2. Are there differences in the extent to which individuals rely on stereotypes when rating males versus females?
3. Do males versus females differ in their levels of stereotype consistency?
4. Do male and female judges differ in stereotype utilization?

Method

Participants

This study used participants from the Idaho State University (ISU) participant pool and from the online work base Amazon Mechanical Turk (MTurk). ISU students were individuals who were taking a class with a research requirement. This is mostly freshman-level psychology classes, but some other psychology classes and a few classes outside of the psychology department may also have been included in the participant pool.

MTurk is an online work base where organizations and individuals can post tasks (called HITs) along with an agreed upon compensation amount. Workers then select tasks and follow the directions to complete them. Anyone over 18 can participate on MTurk, but this study was limited to those who were in the United States, had previously completed over 100 HITs, had above a 60% acceptance rate (their work was rarely rejected), and they were Cloud Research (a company, formerly known as Turk Prime, that helps researchers manage their MTurk HITs) approved workers.

The current study used a few different sets of participants and so demographics and information about power analyses and sample sizes are given in each section.

Measures

Big Five Inventory-2

The BFI-2 (see Appendix A) is a 60-item version of the Big Five Inventory (Soto & John, 2017). It is designed to measure the 5-factor personality trait dimensions of open-mindedness, conscientiousness, extraversion, agreeableness, and negative emotionality. With prior validation research, the Cronbach's alpha reliabilities of each of the 12-item domain scales has ranged from .83 to .91. Short phrases made of basic vocabulary are rated using a Likert scale that ranges from

1 (*disagree strongly*) to 5 (*agree strongly*). The BFI-2 was used with targets as a self-report measure and was given to target acquaintances as a self-report and an other-report measure (more information on this in the target procedures section). The BFI-2 was also used by judges to assess personality perceptions that judges have of targets. Judges also completed this as a self-report measure. With this study's sample, all self-reports were combined and scored, and Cronbach's alpha reliabilities of each of the 12-item domain scales had adequate-to-good internal consistency (Extraversion r = .83, Agreeableness r = .74, Conscientiousness r = .82, Negative Emotionality r = .88, and Open-mindedness r = .80.

General Demographics Questionnaire

The general demographics questionnaire asked about age, ethnicity, race, gender identity, education level, religious affiliation, and marital status.

Procedures

The overall method of this study involved judges watching prerecorded interviews of targets and then rating the targets on the items of the BFI-2. Judge's ratings were then compared against a target's distinctive personality profile, a normative profile, and a profile representing personality stereotypes of the targets' gender. This made it possible to ascertain how distinctively, normatively, and stereotypically a judge rated a target on the items of the BFI-2.

This study used a between-subjects design such that each judge observed and rated targets from one gender group. This was done in order to reduce any possible demand characteristics and to make it possible to obtain reliable estimates of accuracy with judges viewing only six targets[1] (previous research has shown that using more than six targets does not

[1] There was an error (described in the target section) that caused 37 judges to view a target whose data cannot ethically be used in this study, so estimates from these judges are based on only five targets.

result in significant gains in the reliability of judgments; Letzring et al., 2016). Each additional category that is added to the study increases the complexity of analysis and multiplies the number of targets and judges needed to obtain reliable estimates of accuracy (Maas & Hox, 2005). As this is the first study to examine the interaction of stereotypes with accuracy, the analysis and overall methodology have been designed to be relatively simple by limiting targets to only cisgender males and cisgender females.[2]

Stereotype Profile Creation

The first step in understanding how stereotypically a judge is rating a target is to ascertain what the general stereotypes are of a group within the population being studied. To achieve this, a sample of 61 individuals were recruited. Ten participants did not pass 80% of attention checks or did not complete the study. This resulted in a sample size of 51 individuals (M_{age} = 31.1; 59% female, 39% male, less than 2% non-binary; 84% White, 2% Asian, 2% Black or African American, 12% Other or multi-racial) with 67% (*n* = 41) from the ISU psychology department’s undergraduate participant pool and 33% (20) from MTurk. A sample size of 50 was initially set as the target size because it is consistent with what has been used previously to assess favorability of the BFI-2 items (Krzyzaniak, 2020). Once participants signed up, they were given access to the Qualtrics study. Upon clicking the link, participants first read the informed consent and agreed to participate in this study. After agreeing to participate, participants were given a modified version of the BFI-2. This was the same as the BFI-2 other-report used elsewhere in this study with a few modifications. Participants saw the phrase “During this portion of the study, try to place yourself in the mind of your friends, classmates, family members, or those

[2] In order to facilitate readability, the remainder of this document will use the terms male and female to refer to cisgender individuals.

around you. As you answer these questions try to answer as you think they would answer." This statement was followed by another line that said "in general my friends, family members, classmates, and those around me see members of this group as someone who:" after which they saw the BFI-2 but organized so that they could simultaneously rate from 1 (*disagree strongly*) to 5 (*agree strongly*) on how they believe others see White non-Hispanic males and White non-Hispanic females on average on that item. After completing all 60 items, participants were then asked to take the self-report BFI-2, a few other self-report measures,[3] and a short demographics questionnaire, after which they were debriefed and thanked for their participation.

After collecting data from 51 participants, the data were analyzed for Cronbach's alpha reliability, which showed that ratings of genders were highly reliable (*a* = .83) suggesting that no more participants needed to be collected to get consistent results. These scores were used to create a stereotype profile for each gender group that was then correlated with judge ratings and target personality profiles. To do this, ratings of each gender group were averaged by item across participants to create a single stereotype profile for each gender category. This resulted in two profiles, one for each group, that represented how each group was rated on average on items of the BFI-2.

Target Stimuli Creation

This study was originally set up to use 18 targets of each gender category for a total of 36 targets. This was done to provide a wide variety of targets to increase the study's external validity and make it possible to conduct a between-subjects design with judges being assigned to observe and rate six targets from a single gender category. After data collection was complete,

[3] These include the trait-level Positive and Negative Affect Schedule, the Satisfaction with Life Scales, and Ryff's Psychological Well-Being scales. These measures will be used for future research.

analysis revealed that one of the targets was only 16 years old, and so all data collected using this target had to be deleted, leaving 35 targets. Only White, non-Hispanic, cisgender individuals were used for this study. Targets were mostly in their early twenties (*M* = 22.1, range: 18–43) with females being slightly older (*M* = 23.4, *SD* = 7.1) than males (*M* = 20.9, *SD* = 3.3).

Targets were recruited through the ISU participant pool. The description of the study on the ISU recruitment program did not mention anything about race or gender, but targets who were a race other than White or a non-binary gender were able to participate, but their data were not used. Targets signed up for a time slot and met with either the primary investigator or a research assistant in a lab in the psychology building on the ISU campus. The description on SONA informed targets of some of the basics of the study and asked them to have a list prepared with the emails and phone numbers of three to five acquaintances (these include family members, friends, and romantic partners). Targets were asked to pick acquaintances whom they have known for at least 6 months and whom they thought knew them well. Targets were informed in the email about how these emails and phone numbers were used and stored.

The 2019 Novel Coronavirus made it so that meeting in-person, especially in close proximity, was difficult and dangerous. For this reason, all targets were asked to confirm that over the past 2 weeks they had no known contact with anyone who had tested positive for Covid and they had not had any Covid symptoms (which was also required for all researchers). Prior to targets arriving at the lab, a researcher set up a computer and opened up a Qualtrics study. Once targets arrived at the lab, they were asked to put on a mask if they were not currently wearing one (the researcher wore a mask during the full duration of the study's data collection procedures). Targets then took a seat in front of the computer that had been set up for them. The first screen seen by participants said, "Thank you for your participation in this study. Please give your SONA ID number to the researcher and then hit next when they say you can begin." At this point, targets saw a page that contained the first informed consent document (a second informed consent was given at the end of the interview), which notified the target of some of the purposes of the study. Targets were asked to agree to their participation in this study and were informed that they had the ability to confirm the use of their video at the end of the interview (after they knew what information they shared during the recorded interaction). After agreeing to participate in the study, targets completed a self-report version of the BFI-2. Periodically throughout the target's study, simple attention checks were embedded. These attention checks looked similar to other questions but asked targets to select a specific answer. After completing the BFI-2, targets were asked to notify the researcher and then hit next. The next page of the study displayed a list of all the interview questions for targets to review. These questions were adapted from another study (Krzyzaniak, 2020; see appendix D for a copy of these questions) and asked about the target's life, hobbies, and goals. These questions were designed to get the target talking about their life in ways that would elicit cues about various personality traits. Targets were given the chance to review these questions while the researchers set up a camera for the interview process. The researcher set up the camera so the target was visible from mid-chest and up with a blank white wall behind. Targets were asked to notify the researcher when they were ready to begin.

Once the target indicated they were ready, they were asked to try to limit their responses to 1 or 2 minutes and then the researcher began the recording and asked the target each question in turn (and asked follow-up questions if answers were too short). The researcher kept track of time and tactfully moved to the next question after the target had passed the 1-minute mark, but before the 2-minute mark. After all the main questions were asked, the researcher ended the recording and informed the target that they would be asked a second set of questions that were not part of the main study. The same method was used to ask questions about a target's perception of how stereotypes about their race and gender are used, but these questions are not part of this dissertation. After this second set of questions, the researcher again ended the recording.

At this point, targets were asked to return to the Qualtrics study and again confirm the use of their target video in this study now that they knew what they said in the video. After giving this confirmation (no one asked for their videos not to be used), participants completed a second set of self-report measures[4] after which they were instructed to speak with the researcher.

For the final portion of this part of the study, targets were asked to email 3–5 acquaintances who had known them for at least 6 months. A template email was provided on the last page of the Qualtrics study. Targets were asked to copy this template and paste it in a draft email individually to each acquaintance with the lead researcher CC’d. Targets were asked to personalize the email by adding their name and their acquaintances’ names and to click send. Targets were then debriefed and thanked for their participation in the study.

Each acquaintance received an email (see Appendix B for a template of this email) from the target that linked to a Qualtrics study. The first page contained simple information about the study and acquaintances were asked to agree to be a part of this study. The next page contained a single text box where acquaintances input an anonymous ID that was linked to a specific target. After this, acquaintances completed an other-report of the BFI-2 where they rated the target on

[4] These included questions about their perceptions of how often they are stereotyped as a member of their race and gender group and the extent to which they feel others use stereotypes when judging them, along with the trait-level Positive and Negative Affect Schedule, the Satisfaction with Life Scales, and Ryff’s Psychological Well-Being scales. These questions and measures will be used for future research.

each item. Acquaintances ended by completing a few more other-reports[5] and the demographics questionnaire, along with a set of questions asking how long they have known the target and their relationship to them. Acquaintances were then debriefed and thanked for their participation. Periodically throughout the acquaintance's study, simple attention checks were embedded (similar to those mentioned previously) to confirm that only reliable data was being used.[6]

On average, each target was rated by two or three (*M* = 2.4, range: 1–4) acquaintances. Data from acquaintances were combined with the self-report of the respective target to create a single personality profile for each target. This was done by averaging a target's self-report BFI-2 with the average of the target's acquaintance ratings (by averaging a target's acquaintance ratings first and then combining it with the self-report, a composite rating is created in which the target's self-report has the same weight as all their acquaintance reports combined). This target personality profile was then correlated with the target's respective stereotype profile to create a stereotype consistency score. This personality profile was also used as the accuracy criteria to estimate distinctive accuracy of the ratings from judges (this is described more in the analysis section).

After video recordings were collected, the lead researcher edited the videos to make them about 4–6 minutes in length. This is in line with what is typically done in this area of research in order to make cross-comparisons between studies easier (Letzring et al., 2016). In order to reach this time limit, videos were edited so that they began after the researcher asked the questions and

[5] These include the trait-level Positive and Negative Affect Schedule, the Satisfaction with Life Scales, and Ryff's Psychological Well-Being scales. These measures will be used for future research.

[6] There was an initial issue with attention checks and some confusion from acquaintances, leading to an abnormally high error rate. To account for this, standard deviations were plotted for each acquaintance and outliers were not used. The attention checks were corrected mid-study, and so only a few acquaintances were impacted.

ended when the participant finished speaking. If the researcher needed to ask any follow-up questions or prompt more of an answer from targets, the video was edited in order to remove any researcher questions or commentary. All answers from a single target were combined into a single video. During this process, a single still frame of the target was captured and cropped so that it only showed the target's face and minor background. This still frame was used to verify that judges did not recognize any targets that they observed and judged (this is explained more in the judge's procedures).

Judgments

Judges were recruited from the ISU participant pool and from MTurk. In order to get stable estimates in multi-level models, it has been shown that level 2 group sizes of 50 and above typically result in unbiased estimates (Maas & Hox, 2005). Power within multi-level models is determined by a number of interacting variables such as the number of levels, the size of the effect, the intraclass correlations, whether effects of interest are fixed or random, the number of groups, and the number of observations in each group (Kreft & De Leeuw, 1998). This makes it difficult to calculate power in multilevel models, but it is possible to get an approximation by conducting power analysis for a multiple regression with the same number of predictors. In order to verify that 50 judges per group would have enough power to detect the expected effects, G*power (Faul et al., 2007) was used to calculate sample size. The "Linear multiple regression: Fixed model, R^2 deviation from zero" was selected and a medium-small effect size was used (f^2 = .05). Research on stereotype accuracy has found that most stereotypes at the group level have large effects (Jussim et al., 2016), but the proposed research used personality stereotypes at the individual level, and so medium to small effects were expected. A power level of .80 and an alpha of 0.05 were used, and the number of predictors was set at five to account for the four profile types being used in the most complex analysis (normative, distinctive, judge stereotype consistency, target stereotype consistency) and the target gender categories. This resulted in a predicted sample size of 196. For this reason, the final anticipated sample size was set at 200, with 100 assigned to each target gender category.

During data collection, attention checks were periodically analyzed and any failing judges were noted and the total judge count was increased to compensate for this. In all, 247 judges were recruited (187 from ISU and 60 from Mturk). Thirty-two ISU judges were removed, two for technical difficulties (i.e., a power outage and a survey error) that made data unreliable, one for viewing less than four targets, and 29 for not passing 80% of attention checks. Six MTurk participants were removed for not passing 80% of attention checks.[7] There were not significant differences in gender (χ^2 [1] = 0.44, p = .50), race for White vs. non-White judges[8] (χ^2 [1] = 2.28, p = .13), or age (t [51.64][9] = 1.34, p = .19) between those who passed 80% of attention checks and those who did not. This left a total of 209 judges (M_{age} = 27.1; 65.1% female, 33.9% male, less than 1% non-binary; 83.2% White, 4% Asian, 4% Black or African American, 9% Other or multi-racial). One hundred five judges viewed female targets, and 104 viewed male targets.

Judges used either the ISU participant system to sign up for a specific time slot or MTurk to get access to the survey. The description made it clear to judges that they would need to be alone for the duration of the study and that they would need to remove any possible distractions.

[7] This may be significantly less than many individuals familiar with MTurk may be used to seeing, but this study limited who was able to participate, so that only those with a proven track record were able to participate in the study.

[8] All non-White counts were collapsed in order to make sure expected counts were greater than five (Campbell, 2007; Cochran, 1954).

[9] A Welch *t*-test was used because of unequal variances.

ISU participants signed up for a specific time slot to meet with a research assistant over Zoom. Up to six ISU participants were able to simultaneously participate in the study. Upon entering the Zoom room, the researcher introduced themself to the ISU judges and asked them to remove any possible distractions. They then confirmed their SONA numbers and the link to the Qualtrics study was sent to them over the Zoom chat. Regarding MTurk, although it is possible to have participants answer screener questions for a small payment and then contact specific individuals for a follow-up Zoom meeting, this was not done in this study for several reasons: there is a strong possibility of attrition when conducting studies this way, limited funds were available for screener questions, and MTurk participation was limited to those who had a proven track record of active participation in research[10], and so there was less need to have a research assistant available (compared to ISU students who did not have a record of participation).

For ISU participants, the first page of the study contained a video of someone not directly involved in this study introducing individuals to the study and informing them of the basics of what would be required from them during the duration of this study (see Appendix C for a copy of the transcript). This video was removed for the MTurk participants because it mentioned things that were specific to ISU, but commensurate information was contained in the listing for this study on the MTurk website. After the instructional video (or immediately upon entering the study for Mturk participants), judges were asked to give their informed consent to this study. After agreeing to participate, judges completed a self-report version of the BFI-2, after which Qualtrics randomly assigned judges to one of the two target-gender groups and then to one of the three target groups within their assigned gender category (with each target group consisting of

[10] This was done by only recruiting Cloud Research-approved participants who had a high number of completed HITs and a low rejection rate.

six targets). Judges then saw a still image of all six targets and were asked to check a box under the target image if they recognized the targets or to select “I don’t recognize any of them.” If a judge did confirm recognizing a target, they were assigned to another target group within their assigned gender category and once again were asked to confirm that they did not recognize anyone. If they still recognized someone, they were assigned to the last group, and once again were asked to confirm that they did not recognize anyone. If they still recognized someone, they were able to continue with the study, but their data were not used (this was the case for two ISU judges and one MTurk judge).

After confirming or disconfirming recognition of the targets, all judges were then randomly assigned one of their six targets and began watching that target’s video. Progression in the survey was locked until the video had finished. Once the video was complete, judges clicked on “next” and rated the target they just watched on the items of the BFI-2, after which they were shown their next target. This process was repeated until all assigned targets were viewed and rated. All judges then completed a set of self-report measures[11] and a demographics questionnaire. Judges were then shown a debriefing page where they were informed about the nature of the study and were either given credit or payment for their participation. As with other portions of this study, attention checks were included periodically throughout the study. Judges also had a unique set of attention checks after viewing and rating the second and fourth targets. This attention check presented judges with three possible topics, only one of which was mentioned in the target video they watched.

[11] These include the trait-level Positive and Negative Affect Schedule, the Satisfaction with Life Scales, and Ryff’s Psychological Well-Being scales. These measures will be used for future research and were removed for the Mturk portion of the study to save time and resources.

Results

Overall Accuracy

The normative profile was created by averaging together all judge, target, and acquaintance self-reports on the BFI-2. Averages were computed separately for each gender to create normative profiles that represent the average item-level score for that gender ($n_{male} = 132$, $n_{female} = 261$). The purpose of this normative profile was to create a representation of what individuals are like on average. For this reason, self-reports for judges, targets, and acquaintances were used because they would provide a good representation of the population being studied and a strong approximation of what this group is generally like.

The SAM was used to analyze the data which is typically represented using an unstandardized regression equation (see equations 1.1 and 1.2). This equation represents a multi-level model that examines the relationship between a judge's ratings and a target's distinctive profile and the normative profile. This makes it possible to account for the nesting of judges within targets and targets with judges, across multiple items.

$$Y_{jti} = \beta_{0jt} + \beta_{1jt}TCrit_{ti} + \beta_{2jt}Norm_i + \varepsilon_{jti} \qquad 1.1$$

$$\beta_{0jt} = \beta_{00} + u_{0j} + u_{0t} + u_{0(jt)} \qquad 1.2$$

$$\beta_{1jt} = \beta_{10} + u_{1j} + u_{1t} + u_{1(jt)}$$

$$\beta_{2jt} = \beta_{20} + u_{2j} + u_{2t} + u_{2(jt)}$$

Using this model, Y_{jti} represents the estimated accuracy score for judge *j*'s rating of target *t* on item *i* of the BFI-2. $TCrit_{ti}$ is the accuracy criterion (the combined target self-report and acquaintances other-report) of target *t* on item *i*. $Norm_i$ is an estimate of the average rating of all targets of the same gender on item *i*. Prior to analysis, $Norm_i$ was subtracted from $TCrit_{ti}$ and both were mean-centered, making it so that accuracy estimates could be interpreted as the average level of normativity or distinctive accuracy when the other is held constant. β_{0jt} represents the intercept or the expected level of accuracy when $TCrit_{ti}$ and $Norm_i$ are at their mean levels. In this equation, distinctive accuracy is represented by the coefficient β_{1jt}, which represents the estimate of distinctive accuracy when $Norm_i$ is held constant at the mean. Normativity is represented by the coefficient β_{2jt}, which is an estimate of normativity when distinctive accuracy is held constant at the mean.

Within the second level of this equation (1.2), u_{0j}, u_{0t}, and $u_{0(jt)}$ represent, respectively, the random intercepts of the judge, the target, and the judge-target pair. The random effects u_{1j} and u_{2j} represent the random slopes and the residual variance for the judge on distinctive accuracy and normativity, and the random effects u_{1t} and u_{2t} represent the residual variance for the target on distinctive accuracy and normativity, respectively. Lastly, $u_{1(jt)}$ and $u_{2(jt)}$ represent the residual variance for the judge-target pair on distinctive accuracy and normativity. Initial analysis with target, judge, and dyad random effects all failed to converge, and so the dyadic random effect was dropped, which made model convergence possible.

At the basic level (analysis without any moderators, known as the base model), judges on average made accurate judgments of targets' personalities for both distinctive accuracy (*b* = 0.24 [.04][12], , *p* < .001) and normativity (*b* = 0.72 [.05], , *p* < .001). The model was then run using dummy codes with 0 representing the in-group (i.e., if a judge rated targets of the same gender category as themselves) and 1 representing the outgroup (i.e., when judges were a different gender from the targets they rated). For both distinctive accuracy (*b* = 0.03 [.02], , *p* = .16) and normativity (*b* = -0.02 [.05], , *p* = .71), the interaction terms were non-significant, meaning that

[12] Values in brackets represent the standard errors of estimate.

judging targets of the same versus a different gender did not have a significant impact on levels of accuracy[13].

The next step was to create stereotype profiles for each gender so that target accuracy profiles and judge ratings could be compared against these profiles in order to estimate the similarity between the stereotype profile and a target's personality profile or a judge's rating of a target. In order to examine whether there were any meaningful differences between how male and female raters perceived the average White cisgender male or female, stereotype perception data were analyzed by running an independent samples *t*-test for each item of each gender profile, separated by the gender of the rater. This resulted in 120 *t*-tests, and so by chance, it would be expected that six would be statistically significant at a .05 alpha level (which was the case). The sample size for this analysis was also small, with only 20 or 30 raters in each group, and so Cohen's *d*s were calculated in order to look at effect sizes in standard deviation units. In all, for the male stereotype profile, 21 out of the 60 items were above the traditional 0.20 cut-off for a small effect, and eight were above the 0.50 cut-off for a moderate effect (with the largest being 0.67). For the female profile, 29 out of the 60 were above 0.20 and, four were above 0.50 (also with the largest being 0.67). Results were broken down by trait (see Table 1), and the results showed that effects were spread differently across traits, indicating that there were larger differences in how females and males were rated by male and female perceivers on some traits (e.g., agreeableness having the most effects greater than 0.20) compared to others (e.g.,

[13] Accuracy was also examined using self-other agreement, in which the accuracy criterion only used the targets' self-ratings and not the acquaintance ratings, for judges only and for acquaintances only. Self-other agreement was higher for acquaintances on both distinctive accuracy (*b* = 0.36 [.03], *p* < .001) and normativity (*b* = 1.11 [.05], *p* <.001) compared to judges on distinctive accuracy (*b* = 0.16 [.03], *p* < .001) and normativity (*b* = 0.74 [.05], *p* <.001).

Table 1

Stereotype Profile Effect Sizes of Male Versus Female Perceivers of Males and Females

	Ext		Agr		Con		Neg		Ope		Total	
	M	F	M	F	M	F	M	F	M	F	M	F
d = 0.20–0.50	4	5	5	7	7	4	2	6	3	7	21	29
d > 0.50	0	0	3	0	2	1	3	2	0	1	8	4
d > 0.20	4	5	8	7	9	5	5	8	3	8	29	33
Genders combined	9		15		14		13		11		62	

Note. M = Male, F = Female, Ext = Extraversion, Agr = Agreeableness, Con = Conscientiousness, Neg = Neuroticism/ Negative emotionality, Ope = Openness to experiences.

extraversion having the fewest). Only 10% of items overall had a moderate effect size ,and 44% had a small effect size. This means that most effects were small, and only a few were moderate, so it was determined that there were not any meaningful differences in how male and female perceivers rated males or females on the BFI-2.

The next step was to calculate an average stereotype profile for males and females. This was done by averaging the item-level stereotype ratings for males and females across participants. This created an item-level average for each gender on each item of the BFI-2. It is important to note that there is some conceptual overlap between the normative profile and the stereotype profiles. Both profiles represent group-level information and averages for a specific group, but the normative profile was created using the average self-ratings on the BFI-2, while the stereotype profiles were created by asking individuals to rate how they believe their friends, family members, or others would rate non-Hispanic White cisgender males and females. Both scores were designed to represent group-level information and so both profiles represented normative ratings to some extent. The main differences between the profiles come from how they were created. The normative profile represents a more concrete description of what a group is actually like because it was made by averaging a few hundred self-ratings on the BFI-2. Contrastingly, the stereotype profiles nominally represent a more abstract idea of a group because it was created by averaging the perception of how individuals believed their friends, family members, or others would rate a group on average. Additionally, the normative profile was created using individuals from a variety of demographics (with the caveat that it was based on a group that was still predominantly cisgender and White) while the stereotype profile was designed to be based solely on White cisgender individuals. Overall, this means that the normative profile should represent a more concrete profile of what individuals are like on average; whereas, the stereotype profile should represent the abstract perception people have about what a group is like on average. Based on the idea that there is a high degree of accuracy in stereotypes (at least for age and gender), there should be a high degree of overlap between the normative profile and a stereotype profile, but there should also be some degree of differentiation.

To directly test this, the normative profile was correlated with each stereotype profile. Results showed that females were viewed more normatively (r = .82, 95% CI [.72, .89], n = 262, p < .001) than males (r = .44, 95% CI [.21, .62], n = 132, p < .001), and female and male stereotype profiles were not significantly correlated with each other (r = .24, 95% CI [-.01, .46], p = .06). This suggests that for females, there is a strong overlap between the normative profile and perceived personality stereotypes of females. There is also a moderately strong overlap between males and their normative profile. This finding suggests that the relation between normativity and a stereotype profile may be driven by this overlap and not necessarily by the specific stereotypes people hold about a gender (this also suggests that individuals, at least with this sample, may be more accurate in their stereotypes of females than they are of males). For this reason, results for normativity will only be reported for this paper in those analyses where the stereotype profile is not present.

Hypothesis 1

The first hypothesis for this project states that a judge's level of distinctive accuracy will be moderated by a target's level of stereotype consistency and by a judge's utilization of stereotype information. If this is true, results should show that judges who make more stereotype-consistent judgments will be more distinctively accurate when targets' personalities are consistent with their stereotype profile, and judges who make less stereotype-consistent judgments will be more distinctively accurate when targets' personalities are inconsistent with their stereotype profile.

In order to test this hypothesis, all targets' personality profiles (i.e., the combined target self-rating and acquaintance ratings) were first correlated with the target's corresponding gender stereotype profile (created by averaging the male or female ratings from the stereotype profile). This provided a profile correlation coefficient, referred to as *target stereotype consistency*, for each target that represents how consistent each target's personality is with their respective stereotype profile. On average, targets had medium to high correlations (Cohen, 2013; Gignac & Szodorai, 2016) with their respective gender stereotype profile (M_r = .49, *SD* = .19, range: -.06–.79). In addition, all judge ratings of an individual target were correlated with that target's corresponding gender stereotype profile. This provided a correlation coefficient for each judge-target pair that represented how stereotypically a judge rated a specific target. This correlation is referred to as *judge stereotype consistency*. Judge ratings on average had small to medium correlations with the stereotype profile (M_r = .27, *SD* = 28, range: -.60 – .82). These correlations were then mean-centered and transformed into *z*-scores using Fisher's transformation.

To test the assumption that the data are normally distributed, the Jarque-Bera normality test (Jarque & Bera, 1980) was used on both judge and target stereotype consistency. Judge stereotype consistency was not normally distributed (*JB* = 53.73, *p* < .01) with a negative skew, while target stereotype consistency was normally distributed (*JB* = 0.16, *p* = .92). All SAM analyses were computed with and without transforming judge stereotype consistency (by multiplying it by -1 to reflect it because it was negatively skewed, adding 10 to remove all negative values, and then taking the log10 of the values); this resulted in a slightly more normal distribution (*JB* = 19.36, *p* < .001). Results using the original and transformed variables were visually compared by the lead researcher, and in all cases, there were only small differences between the estimates. Only in one instance was there was a significant difference, but this difference did not change the interpretation of results and will be noted where applicable. Interpreting estimates from transformed variables is not straightforward (especially when dealing with the complexity of multi-level models), and research suggests that multi-level models are extremely robust to predictor variables that violate normality, and in almost all cases (unless variables are bimodal, which was not the case) there is no significant impact on estimates (Maas & Hox, 2004; Schielzeth et al., 2020). For these reasons, the non-transformed variable was used to make the interpretation of the results more straightforward and easier to understand. Although predictor variables can be non-normal, error terms need to be normally distributed, and so Q-Q plots were used to test this in all analyses. In all models, residuals were normally distributed, indicating that estimates can be relied on to make predictions.

Hypothesis 1 can be represented in the following standardized regression equation:

$$Y_{jti} = \beta_{0jt} + \beta_{1jt}TCrit_{ti} + \beta_{2jt}Norm_{i\,+}\,\varepsilon_{jti}$$

$$\beta_{0jt} = \beta_{00} + \beta_{01}\textit{T-Stereotype consistency} + \beta_{02}\textit{J-Stereotype consistency} +$$

$$\beta_{03}\textit{T-Stereotype consistency* J-Stereotype consistency} + u_{0j} + u_{0t} + u_{0(jt)}$$

$$\beta_{1jt} = \beta_{10} + \beta_{01}\textit{T-Stereotype consistency} + \beta_{12}\textit{J-Stereotype consistency} +$$

$$\beta_{13}\textit{T-Stereotype consistency* J-Stereotype consistency} + u_{1j} + u_{1t} + u_{1(jt)}$$

$$\beta_{2jt}= \beta_{20} + \beta_{21}\textit{T-Stereotype consistency} + \beta_{22}\textit{J-Stereotype consistency} +$$

$$\beta_{23}\textit{T-Stereotype consistency* J-Stereotype consistency} + u_{2j} + u_{2t} + u_{2(jt)}$$

Within this model, the moderator *T-Stereotype consistency* represents target *t*'s *z*-transformed correlation with their respective gender stereotype profile. This correlation was mean-centered, so higher numbers represent a greater correlation on average (compared to other targets) between a target's personality profile and their stereotype profile. If this main effect is positive and significant, it means that, on average, the stronger and more positive the correlation is between a target and their respective stereotype profile, the more normatively or distinctively accurate targets were rated by judges. The moderator *J-Stereotype consistency* represents the *z*-transformed correlation between judge *j*'s rating of target *t,* and target *t*'s respective stereotype profile. This correlation was also mean-centered, and so higher numbers represent a greater correlation on average (when compared to other ratings of targets) between all judges' ratings and the targets' respective stereotype profiles. If this main effect is positive and significant, it means that, on average, how stereotypically a judge rated a target is related to how normative or distinctively accurate judges are in their ratings of targets. Hypothesis 1 specifically predicted an interaction between how stereotypically a judge rates a target and that target's own correlation to the stereotype. This is represented by the interaction term *T-Stereotype consistency*J-Stereotype consistency*. If the interaction terms are statistically significant, it can be interpreted to mean that, on average, these terms interact to predict normativity and distinctive accuracy. A positive coefficient means that judges who rate targets more stereotypically when targets are actually more stereotypical, and judges who rate targets less stereotypically when targets are actually less stereotypical, achieve higher levels of normativity or distinctive accuracy. A negative coefficient means that judges who rate targets more stereotypically when targets are less stereotypical, and judges who rate targets less stereotypically when targets are actually more stereotypical, achieve higher levels of normativity or distinctive accuracy. Essentially this would suggest that when judges are sensitive to how stereotypical a target is, and they make more stereotypical judgments when targets are, in actuality, more stereotypical, but make less stereotypical judgments when targets are less stereotypical, then normativity or distinctive accuracy would be higher.

To examine this, J-Stereotype consistency was first added to the model in order to examine the extent to which judges' stereotype use, apart from targets' stereotype consistency, moderated accuracy. There was a significant interaction between J-Stereotype consistency and distinctive accuracy (b = 0.09 [.01], $p < .001$). This model was compared to the base model in a one-way ANOVA as a test of the main effects, which showed that the larger model with the interactions accounted for a significant amount of additional variance (χ^2 [3] = 911.47, $p < .001$). To directly examine the interaction, J-Stereotype consistency was recentered either 1 standard deviation above (i.e., high condition) or one standard deviation below (i.e., low condition) the mean, and then the analyses were computed again (see Latif et al. [2022] for an example of this procedure). Doing this recalculates accuracy estimates relative to the center of each moderating variable so that interaction terms can be interpreted to estimate the level of accuracy for high or low levels of each variable. This showed that, on average, judges with high stereotype consistency (judges who rated targets more stereotypically) were less distinctively accurate (*b* = 0.15 [.004], *p* < .001) than judges with low stereotype consistency (*b* = 0.33 [.04], *p* < .001), but in both cases, ratings were statistically significant, indicating that both groups were still accurate in their ratings of targets.

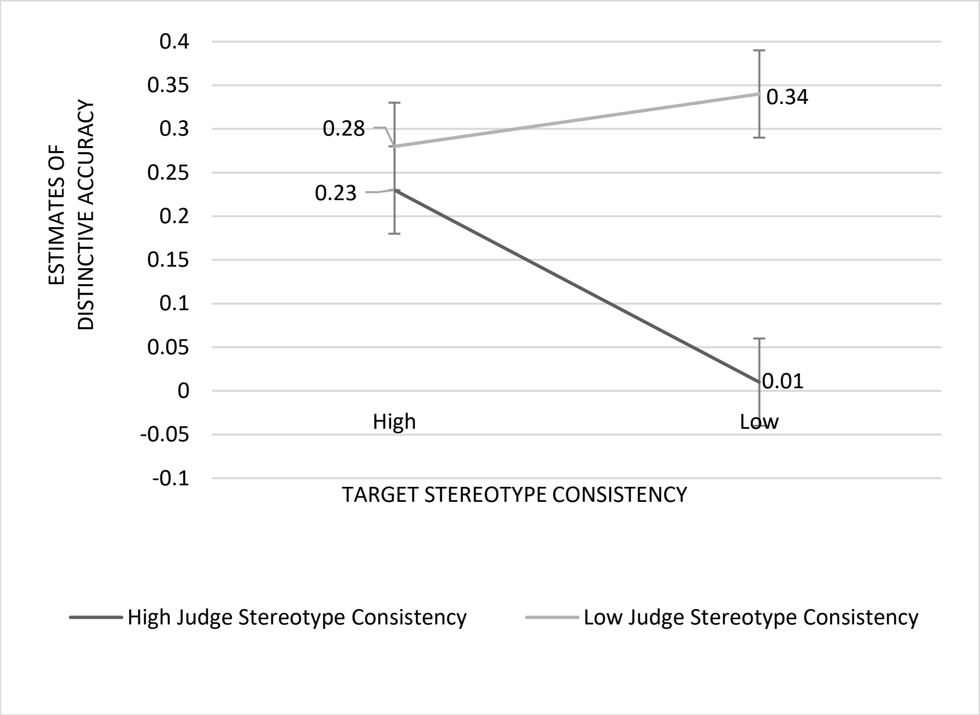
Next, T-Stereotype consistency was added to the model to create a 3-way interaction with J-Stereotype consistency and accuracy. The 3-way interaction was significant for distinctive accuracy (*b* = 0.07 [.006], *p* < .001), indicating that judge stereotype consistency interacted with target stereotype consistency in predicting distinctive accuracy. This model was compared to the previous model with only J-stereotype consistency in a 1-way ANOVA, which showed that the larger model with the additional interaction of T-Stereotype consistency accounted for a significant amount of additional variance (χ^2 [6] = 129.12, *p* < .001). To see how target stereotype consistency individually interacted with accuracy while holding judge stereotype consistency at the mean, accuracy was computed by centering T-Stereotype consistency 1 standard deviation above (i.e., high condition) and below (i.e., low condition) the mean. Results indicated that when targets were more stereotypical, they were rated with less distinctive accuracy (*b* = 0.03 [.009], *p* = .002) compared to less stereotypical targets (*b* = 0.16 [.01], *p* < .001). In order to further describe the interaction, four models were computed by centering J-Stereotype consistency 1 standard deviation above (high condition) and below (low condition) the mean, and by centering T-Stereotype consistency 1 standard deviation above (high condition) and below (low condition) the mean (see Figure 4). When judges' ratings were high in stereotype consistency (or highly stereotypical in their perceptions of targets), targets who had high stereotype consistency (or who were highly stereotypical) were judged with greater distinctive accuracy (*b* = 0.23 [.05], *p* < .001) compared to targets who were less stereotypical (*b* = 0.01 [.05], *p* = .77; in this case judgments were no longer statistically significant)[14]. When judge ratings were low in stereotype consistency, they were less distinctively accurate when making judgments of targets with high stereotype consistency (*b* = 0.28 [.05], *p* < .001) but more distinctively accurate when making judgments of targets low on stereotype consistency (*b* = 0.34 [.05], *p* < .001).

Judge stereotype consistency scores are somewhat independent across targets, meaning that judges could have high stereotype consistency for some targets and low consistency for others. This could be a driving force behind the results presented so far, but they could also be driven by judges who, on average, are high or low on stereotype consistency across targets. To further pull apart the driving forces behind this finding, judges' stereotype consistency scores were averaged across targets to create a single score representing a judge's average level of stereotype consistency across targets. This variable was used to predict distinctive accuracy. The results indicated that there was a significant interaction with distinctive accuracy (*b* = 0.03 [.01], *p* = .001). As before, this variable was recentered either one standard deviation above (high condition) and below (low condition) the mean, and judges who made on average more stereotypical judgments were less accurate (*b* = 0.21 [.04], *p* < .001) compared to those who made less stereotypical judgments (*b* = 0.27 [.04], *p* < .001).

[14] This is significant if the transformed J-stereotype variable is used.

Figure 4

Distinctive Accuracy for High or Low Levels of Judge and Target Stereotype Consistency



Note. Error bars represent Standard Errors.

Hypothesis 2

The second hypothesis states that judges with greater favorability, or those who have a more positive perception, of a specific group, would be less likely to ignore individuating information about specific targets (possibly due to less confirmation bias) and therefore would judge targets from that group with greater variability and higher levels of distinctive accuracy.

To test this, judge’s ratings of targets were averaged at the item level to create a profile of mean ratings for each judge across all six targets, and then this rating was correlated with favorability ratings of the BFI-2 items[15]. Other researchers collected the favorability ratings as part of a separate project (Krzyzaniak, 2020).[16] The judge's average target ratings were correlated with this favorability profile. Higher correlations indicate that judges viewed this group on average in a more favorable way. This is referred to as *J-favorability*. This method ascertained how favorably a judge viewed a specific group without having to directly ask them about favorability, which reduces demand characteristics. Hypothesis 2 can be represented in this standardized regression equation:

$$Y_{jti} = \beta_{0jt} + \beta_{1jt}TCrit_{ti} + \beta_{2jt}Norm_{i\ +}\ \varepsilon_{jti}$$

$$\beta_{0jt} = \beta_{00} + \beta_{01}J\text{-}favorability + u_{0j} + u_{0t} + u_{0(jt)}$$

$$\beta_{1jt} = \beta_{10} + \beta_{11}J\text{-}favorability + u_{1j} + u_{1t} + u_{1(jt)}$$

$$\beta_{2jt}= \beta_{20} + \beta_{21}J\text{-}favorability + u_{2j} + u_{2t} + u_{2(jt)}$$

As mentioned, the moderator *J-favorability* represents the correlation between judge *j*'s perceptions of their assigned gender group and a favorability profile. This variable was mean-centered, so a positive value means the ratings are more favorable than the average. A positive

[15] In the original study design, favorability was intended to be assessed by having judges complete an other-report of the BFI-2 about how they see White males or females on average. Judges who observed and rated males would respond to the questionnaire about males on average, and judges who observed and rated females would respond about females on average. Unfortunately, an extra word (females) was added after the dynamic code in Qualtrics that would instruct the judges to answer the question about males or females depending on the judge's group. Because of this, many judges instead saw the phrase "… I see white females females …," or "… I see white males females …" This error was noticed halfway through data collection and because it made the data meaningless, a different approach was use to determine favorability.

[16] Individuals from the ISU participant pool were asked to rate each item of the BFI-2 (along with the BFI and the NEO-PI) on a 5-point scale on how favorable or likable they would find a person who fits that description. Fifty-two participants were originally recruited with the goal of reaching a Cronbach's alpha of .80 across items of the BFI-2. Any participants who did not pass 80% of the embedded attention checks were dropped, leaving a total of 38 participants with an alpha of .81.

and significant main effect can be interpreted to mean that, on average, the more favorably a judge rates their target group, the higher their predicted normative or distinctive accuracy.

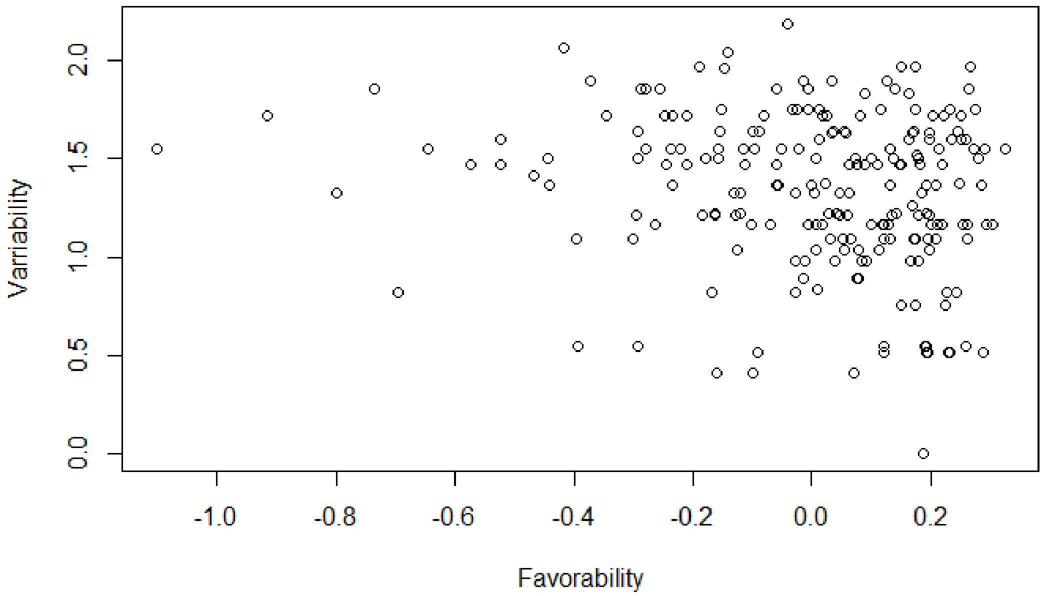
Further support for this hypothesis was investigated by looking at the variability of ratings. Judges with more negative views were predicted to have less variability in their ratings of targets compared to judges with more positive views because judges with negative views were assumed to be more likely to rely on their own stereotypes (which would be stable) and less on the individuating information of the target (which is more variable across targets). This was tested by calculating the variance for each judge's ratings of all of the targets, for each item, and correlating that with favorability. A significant positive correlation would suggest a relation between a judge's evaluation of a group and the variability of judge's ratings of targets (with more favorability leading to more variability in judgments).

Favorability did not significantly moderate distinctive accuracy (b = 0.04 [.04], p = .36), but did significantly moderate normativity (b = 1.20 [.07], p < .001). This model was compared to the base model in a one-way ANOVA, which showed that the larger model with the additional interaction of favorability accounted for a significant amount of additional variance (χ^2 [3] = 200.13, p < .001). When judge variability (calculated as the average item-level standard deviation for each judge) was correlated with judge favorability, results indicated a significant but small negative correlation (r = -.17, 95% CI[-.30, -.04], p = .01) indicating that variability correlated negatively to favorability to a small significant degree (see Figure 5).

An independent samples t-test was computed to investigate differences in how favorably judges viewed male versus female targets. There was a significant difference in favorability between the groups (*t* [207] = 2.51, *p* = .01, *d* = 0.34)[17], with male targets being viewed more favorably (*M* = .04, 95% CI [.04, .05], *SD* = .26) than females (*M* = -.04 [-.043, -.035], *SD* = .21).

Figure 5

Judge Favorability by Judge Variability



Research Questions

1. Are there any differences in normativity or distinctive accuracy for male versus female targets?

In order to test the first question, the base model was slightly altered to examine how accuracy levels differed by gender of targets.

$$Y_{jti} = \beta_{0jt} + \beta_{1jt}Tcrit_{ti} + \beta_{2jt}Norm_{i} {}_{+} \varepsilon_{jti}$$

$$\beta_{0jt} = \beta_{00} + \beta_{01}G + u_{0j} + u_{0t} + u_{0(jt)}$$

[17] These results were computed using the mean-centered variable and should be interpreted accordingly.

$$\beta_{1jt} = \beta_{10} + \beta_{11}G + u_{1j} + u_{1t} + u_{1(jt)}$$

$$\beta_{2jt} = \beta_{20} + \beta_{21}G + u_{2j} + u_{2t} + u_{2(jt)}$$

This model used dummy coding to represent the groups under investigation. Within this model, the dummy code is used to represent whether targets were female (coded as 0) or male (coded as 1). This means that all output from this model will be relative to the comparison group (i.e., White females), with positive numbers representing more normativity or distinctive accuracy and negative numbers representing less normativity or distinctive accuracy, compared to judging White males. Both distinctive accuracy (*b* = 0.07 [.08], *p* = .34) and normativity (*b* = -0.07 [.09], *p* = .43) did not significantly differ by group, suggesting that whether targets were male or female did not significantly impact accuracy.

2. Are there any differences in the extent to which individuals rely on stereotypes when rating males versus females?

To test this, an average *J-stereotype consistency* score was calculated for each judge and separated by gender of targets, and an independent samples *t*-tests was run. When looking at these results, and all other results of the research questions, it is important to remember that these numbers are based on the mean-centered and *z*-transformed variables, and so 0 represents the mean of both groups combined with negative and positive numbers indicating the distance from the mean. There was a significant difference (*t* [207] = -7.73, *p* < .001, *d* = 1.07), with females being rated with greater stereotype consistency (*M* = .47 , 95% CI[.45, .49], *SD* = .87) than males (*M* = -.47, 95% CI [-.49, -.46], *SD* = .90; see Figure 6).

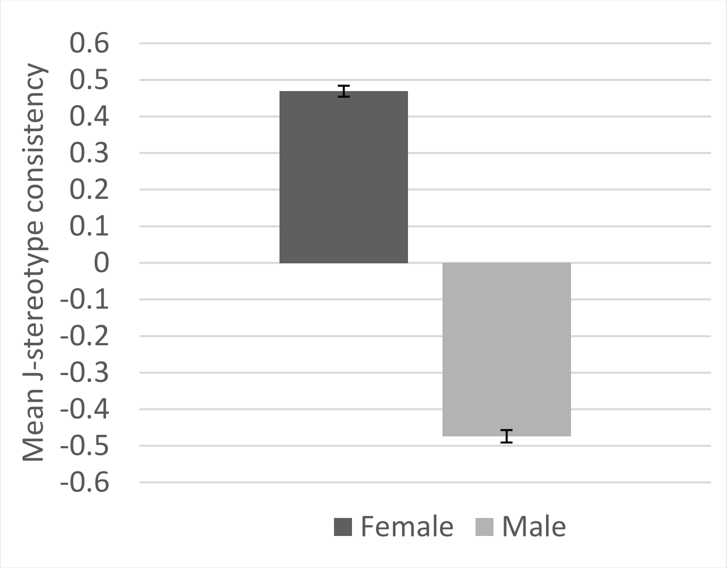
3. Do male and female targets differ in their levels of stereotype consistency?

To test this, an independent samples *t*-test was run comparing each group's *T-Stereotype Consistency* scores. There was not a significant difference, but the effect size was moderate (*t*

[33] = -1.78, *p* = .08, *g* = 0.59[18]) between males (*M* = -.31, 95% CI [-.42, -.19], *SD* = .97) and females (*M* = .29, 95% CI [.18, .40], *SD* = .01; see Figure 7).

Figure 6

Mean J-Stereotype Consistency Score by Target Gender

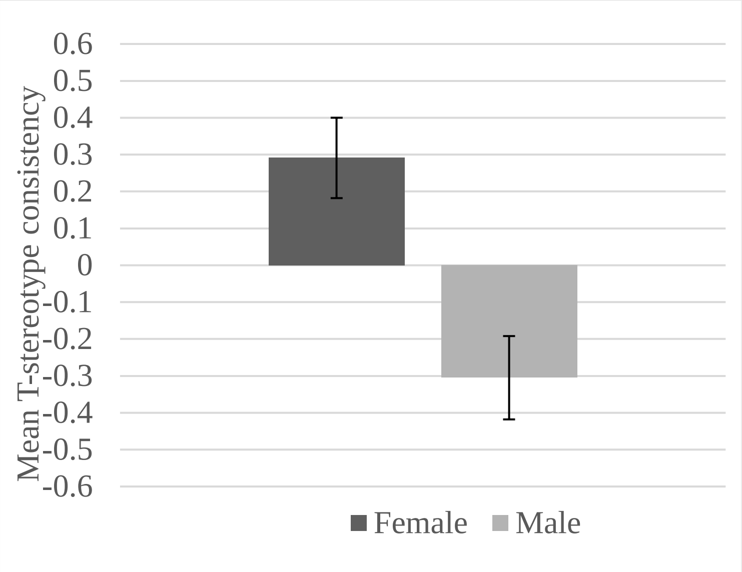


Note. Error bars represent the 95% confidence intervals. Scores were mean-centered and *z*-transformed prior to being separated by gender.

[18] Hedge's correction was used because of the small sample size.

Figure 7

Mean T-Stereotype Consistency Score by Gender



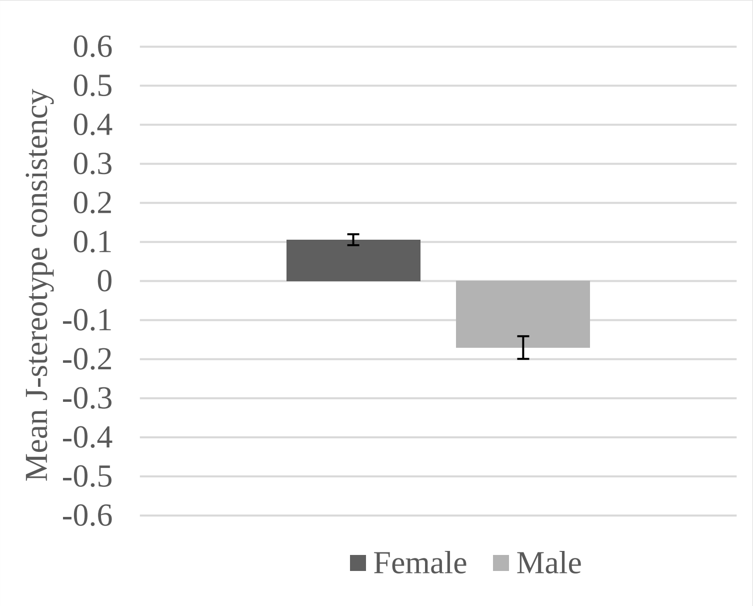
Note: Error bars represent the 95% confidence intervals. Scores were mean-centered and *z*-transformed prior to being separated by gender.

4. Did male and female judges differ in stereotype utilization?

This was tested by an independent samples *t*-test comparing each judge's average *J-Stereotype consistency* between male and female judges. There was not a significant difference (though results did approach significance) and only a small effect (*t* [205] = -1.90, *p* = .06, *d* = 0.28), with males having lower average *J-Stereotype consistency* (*M* = -.17, 95% CI [-.20, -.14], *SD* = .04) compared to females (*M* = .01, 95% CI [.09, .12], *SD* = .97; see Figure 8).

Figure 8

Mean J-Stereotype Consistency by Judge Gender



Note: Error bars represent the 95% confidence intervals. Scores were mean-centered and *z*-transformed prior to being separated by gender.

In order to better understand how gender is related to stereotype consistency, a two-way ANOVA was used to look at the interaction of judge gender and target gender in predicting a judge's average stereotype consistency. Results indicated that there was not a significant interaction (*F* [1, 1] = 0.34, *p* = .56), indicating that male judge stereotype consistency between male and female targets was not meaningfully different than female judge stereotype consistency between male and female targets.

Discussion

Overall results for this study were mixed but still important and interesting. On average, all judges made significantly accurate distinctive and normative judgments of targets. One of the main purposes of this project was to address some of the unasked questions in relation to stereotypes and stereotype accuracy (Jussim et al., 2018). Two of these questions were tested in Hypothesis 1 and 2 respectively.

Hypothesis 1

“When and how does relying on a stereotype increase the accuracy of person perception?” (Jussim et al., 2018, p. 214)

Hypothesis 1 predicted that a judge’s level of distinctive accuracy would be moderated by a target’s level of stereotype consistency and by a judge’s utilization of stereotype information. This was supported by the finding that a judge’s level of stereotype consistency significantly interacted with a target’s level of stereotype consistency in predicting distinctive accuracy. Hypothesis 1 further predicted that judges who make more stereotype-consistent judgments would be more distinctively accurate when targets’ personalities are more consistent with their stereotype profile than when targets’ personalities are less consistent with their stereotype profile, and judges who make less stereotype-consistent judgments would be more distinctively accurate when targets’ personalities are less consistent with their stereotype profile than when targets’ personalities are more consistent with their stereotype profile. This was also supported by the finding that distinctive accuracy estimates were higher when there was a correspondence between how highly stereotypical a target was and how highly stereotypical a judge rated the target. In other words, judges who made stereotypical judgments of a target were more distinctively accurate when targets were highly stereotypical but less accurate when targets were not stereotypical. The reverse was true when judges made judgments that were not highly stereotypical.

This finding is interesting because it adds important caveats to arguments that stereotypes should never be used because they are unfair to the individual being judged (Fiske, 1989; Stangor, 1995, 2016) or that they only hold a small kernel of truth and are generally inaccurate (Ellemers, 2018). Consistent with previous research (Jussim et al., 2021; Löckenhoff et al., 2014), the personalities of individuals on average were moderately correlated with their stereotype profile, suggesting more than a kernel of truth (though never a perfect representation) to the stereotypes of the group to which they belong. This indicates that stereotypes may have utility in their ability to act as a heuristic when making judgments of personality but only in situations when judges make stereotype consistent judgments of targets who are stereotypical. Overall, making less stereotype-consistent judgments led to better accuracy. This means that in every day interactions, judges who make less stereotype consistent judgments will make more accurate judgments on average (at least when it comes to White non-Hispanic cisgender college-age individuals). This suggests that although personality stereotypes tend to have moderate to high levels of accuracy, their utility at the group level may be limited to the specific set of circumstances just mentioned.

It is important to note here that the largest differences in accuracy were found among targets with low stereotype consistency (see Figure 4). This means that if a target is not highly stereotypical, they will be judged with much greater accuracy when judges rely less on stereotypes but accuracy will greatly decrease if judges are unable to disregard stereotype information. This has important implications for many aspects of life such as dating, getting a job, or even daily interactions. Targets low in stereotype consistency may have to deal with greater variability in the judgments individuals make of them compared to targets who are high in stereotype consistency.

When looking at this from the judge's side, when judges made less stereotype-consistent judgments, they were overall more accurate compared to when judges made more stereotype-consistent judgments. This finding has several potential implications. At face value, it can be interpreted to mean that if judges use stereotypes less, they will be overall more accurate in their judgments. It may be that relying on stereotypes causes judges to be less sensitive to individuating information and therefore fail to detect, or potentially incorrectly utilize, personality cues from the target. Alternatively, it may not be the stereotypes themselves that lead to more or less accuracy. This study did not investigate causality, and so it is possible that a third variable, such as judge’s ability to use individuating information, could cause judges to both rely on stereotypes less and make more distinctively accurate judgments. It may be that one feature of the good judge is that they intuitively know when to rely on stereotypes and when to disregard them. It may be that if a judge who is not as sensitive to individuating information is asked to rely less on their stereotypes, they would have less accuracy in their perceptions because they are unable to compensate for this with attention to individuating personality cues.

The RAM conceptualizes accuracy as the product of a target who generates personality cues that are relevant and available to the trait being judged, and a judge who correctly detects and utilizes these cues. Within the framework of RAM, stereotypes would act as a type of prior information (Gibson, 2019) and can also be cues themselves. This may mean that a judge’s level of stereotype consistency could moderate either the detection of these cues or how judges utilize them. This may lead to confirmation bias which causes judges to skew what cues they are sensitive to (detection) and/or how they interpret cues (utilization). This may be related to more accurate judgments when a target is stereotypical but to decreased accuracy if the target is not stereotypical, which is exactly what was found in the current study.

The RAM also describes four moderators of accuracy (good: judge, target, trait, and information). Results from this study may suggest that either good judges are those who can flexibly use stereotypes when they apply to a target and focus more on individuating information if they do not, or that good judges are those who in general use stereotypes less. Either explanation may help explain why accuracy was higher when judges were less stereotype-consistent. This study also demonstrated a connection between the "good information" moderator of RAM and stereotypes by showing that the quality of the information being received can directly impact how cues are detected (by altering what cues they attend to) and utilized (by altering how they interpret those cues).

Future research should test the causality of these findings by directing judges to refrain from using stereotypes when making judgments or presenting them with purposefully true and false stereotypes in order to directly manipulate prior information and cue relevance. It is also important to test this in other populations and see if these findings still hold. This study only looked at White non-Hispanic cisgender targets, but people may hold stronger stereotypes of marginalized groups and so this could impact the replicability of these findings depending on the level of stereotype consistency of group members.

Hypothesis 2

"Do people ever actually ignore individuals' personal characteristics when perceiving, evaluating, and judging them?" (Jussim et al., 2018, p. 214)

Hypothesis 2 predicted that judges with greater favorability would be less likely to ignore individuating information about specific targets and therefore would judge targets from that group with greater variability and higher levels of distinctive accuracy. Results for this hypothesis indicate that a judge’s level of favorability towards a specific gender did not predict how accurately they judged targets. When judge favorability was entered into the model, it did not moderate distinctive accuracy or normativity. This was further tested by creating an average variability score for each judge (by calculating the average item level standard deviation across targets) and comparing that to judge favorability. No relation was found, indicating that how favorably a judge rated a target was not related to the variability with which they rated targets. Follow-up analysis indicated that judges did indeed view female targets with less favorability compared to male targets, but this was not significantly related to accuracy.

This is an interesting finding because it suggests that the valence an individual attached to a group is not related to (or at least has little to no relation with) the variability or accuracy of judges’ ratings–at least in the case of gender among primarily binary-cisgender, non-Hispanic White judges and targets. It is possible that this finding was non-significant because the predominantly White group of judges in this study hold a mostly favorable view of the all-White group of targets, but future research will be needed to examine this possibility by comparing favorability and accuracy between White and not White groups.

It is interesting that even though there were differences in how favorably individuals viewed males versus females (with males being viewed more favorably), this did not seem to make much of an impact on accuracy. The effect size for this was small, and so it may be that a larger sample size would better be able to detect this effect if it does exist. This is even more interesting because it goes against other research that found that on average females were viewed more favorably (Chan et al., 2011). This may be due to differences in how this study versus previous research operationalized favorability (i.e., this study used judges’ ratings of targets compared to a favorability profile while previous research used normativity which has been shown to measure positivity or favorability). Within the framework of RAM, it seems that the favorability a judge holds towards a specific group does not meaningfully impact the detection or utilization of cues, at least when predominately White non-Hispanic cisgender college-age adults make judgments of White non-Hispanic cisgender, predominately college-age targets. Future research can directly test this by manipulating favorability and looking at how this interacts with accuracy and stereotype consistency.

Targets for this study were all White cisgender individuals who may be seen more favorably than other groups, so future research needs to directly test this. Individuals may hold much stronger views of other non-White groups, which could lead to a more noticeable effect on how judges detect and utilize cues. Research has found that negative information is held to more tightly and is harder to retroactively change (Fiske, 1980; Ybarra et al., 1999), and so if an individual holds a strongly negative view of a group they may be more likely to hold to their stereotypes, or if they have a strongly positive view they may be more willing to look past their stereotypes or to use them correctly and this could directly impact this aspect of the RAM. It will be interesting to compare accuracy and stereotype use between different groups and see what impacts outwardly expressed demographics such as race or gender, or typically less obvious demographics such as religion or political ideology, have on stereotype use and accuracy.

Research Questions

This study included several exploratory analyses that did not include any a priori predictions. This means that all research question results should be interpreted with caution. It is also important to remember that all analysis were with White non-Hispanic cisgender, predominately college-age targets and the majority of judges were also White non-Hispanic cisgender college-age adults. Regardless, these findings can act as a catalyst for important and interesting future directions.

1. Accuracy for Male Versus Female Targets

The first research question investigated whether there were any differences in normativity or distinctive accuracy for male versus female targets. Results indicated that a target's gender did not significantly moderate accuracy, suggesting that a target being male or female did not strongly impact how accurately a judge rated them. There has not been a lot published on this idea, but what research has been done generally finds only small differences in accuracy between targets of different genders (Chan et al., 2011; Mignault & Human, 2021) which is consistent with what was found here. It seems that although there is a lot of variability with gender in personality expression, on average males and females exhibit an equivalent quantity and quality of cues that judges can detect and utilize. It seems that regardless of gender (at least with this demographically homogenous sample), individuals are typically equal in their levels of expressiveness (see Mignault and Human [2021] for a short discussion of this idea).

2 & 3. J-Stereotype and T-Stereotype Consistency of Male and Female Targets

The second question investigated the extent to which judges rely on stereotypes when rating males versus females, and the third question investigated the extent to which male versus female targets differ in their levels of stereotype consistency. On average, females were rated as being more stereotypical than males, but the results from the first research question suggest that this did not significantly impact accuracy. For targets, there was a moderate but non-significant difference (which is at least in part due to the small sample size) in stereotype consistency between male and female targets. These results are particularly interesting when taken together because they suggest that although judges rated male and female targets differently with a moderate effect, this did not have any noticeable impact on how accurately individuals were viewed. This is even more intriguing given the large effect size found for question 2.

Future research will be needed to further pull this idea apart. What is it that causes individuals to rate females more stereotypically than males, and what causes this to not have a noticeable impact on accuracy? As with question one, it would be interesting to manipulate the stereotypes people use to see if getting individuals to use more or less extreme stereotypes impacts their accuracy. Is there a point where stereotypes are extreme enough that they significantly impact accuracy?

4. J-Stereotype Consistency for Male versus Female Judges

The final question investigated whether male and female judges differ in stereotype utilization. Question 2 examined the difference in stereotype consistency for male versus female targets, and this question is about the differences in stereotype consistency for male versus female judges. There was a small and non-significant effect, suggesting that males and females did not significantly differ in how much they used stereotypes when making judgments of individuals. Taken with the results of the other research questions, this means that even though judges as a whole differed in how stereotypically they rated male and female targets, this effect was not found to be different for male or female judges. This relation was further investigated in a two-way ANOVA which shows that indeed male judge stereotype consistency between male and female targets was not meaningfully different than female judge stereotype consistency between male and female targets. Previous research on accuracy has found mixed results when looking at how judge gender moderates accuracy (Chan et al., 2011; Colman, 2021), and this may be true for gender-related stereotype use as well. It may be that gender is too broad a category to make such large generalizations about stereotype use, or that there really are no meaningful differences in male and female judges' ability to make accurate judgments, or that there is another cause altogether, and so future studies should investigate smaller groups and see how membership in these groups impact the stereotype consistency of judgments.

Other Implications and Limitations

As was stated earlier, it is important to keep in mind that both stereotypes and normativity are a form of group-level information and do have some overlap. The stereotype profile represents how individuals perceive that their friends, family, or others would rate this specific group on average, while the normative profile represents the actual self-ratings of individuals on average on those items. Results of a correlation analysis found that for females, and to some degree with males as well, there was a high correspondence between normativity and the stereotype profile. Both of these profiles represent information at the group level, which further supports the nomothetic accuracy of stereotypes because it shows that the perceptions people have about a gender stereotype (or how individuals believe others would rate this group on average) have a high degree of overlap with how that group rated themselves on average (keeping in mind that the normative profile was based on ratings of predominantly White and cisgender participants, although there were a few non-binary and non-White individuals as well). Future research should investigate the impact of stereotypes that have a lower degree of accuracy to see what impact this would have. For example, research has found that national character stereotypes (beliefs about how those from different nationalities behave on average) do not have a high degree of stereotype accuracy (Jussim et al., 2021). If the findings from this study hold, then using stereotypes with less accuracy (or stereotype profiles that have a low correlation with some criterion such as the normative profile created from a specific group) will likely lead to even lower levels of accuracy for judgments of individuals. In this case, there might be a larger difference in accuracy between judges and targets who are high on stereotype consistency and those who are low on stereotype consistency, with high stereotype consistency leading to even less accuracy (for both targets and judges).

By far the largest limitation of this study is that it was limited to only White cisgender non-Hispanic individuals. The results were interesting, but it is difficult to know if these results would be the same or more or less extreme with other groups. How salient stereotypes are with other demographic groups may impact these results and so future research needs to look at how these findings extrapolate to other groups. For example, with all research questions, non-cisgender judges were dropped from the analysis because they were such a small subsample (2% or less in all samples), and all targets were cisgender. This is one area that needs significantly more research because there is a lot that is still unknown about these populations. To date, no research on personality judgment accuracy has directly been done with these populations, and so it is unclear how findings will generalize. Specifically, within the context of this study, it is unclear if the typical finding of gender not moderating accuracy would also be found among non-cisgender individuals. Would this group as a whole have higher or lower levels of accuracy? Members of these groups often face adversity and so they may, by necessity, need to be more accurate in their judgments. It is also an open question if non-cisgender individuals would be judged as more or less stereotype consistent and this could impact how accurately they are judged. Future research can investigate if this is moderated by a judge's knowledge of a non-cisgender target's gender identity, or if non-cisgender individuals make more or less stereotype-consistent judgments.

Future research will also be needed to see if the results of this study generalize to other demographic groups. When people think of stereotypes, they often think of marginalized groups, and so it would be important to see if results are more or less extreme when including marginalized groups. If the stereotypes of marginalized groups are more salient, it would make sense that judgments of them could be more consistent with the stereotype of their group, but it is unknown what impact this might have on accuracy. Stereotype accuracy research suggests that overall stereotypes about race likely have moderate or high levels of accuracy (Jussim et al., 2016), and so results would be expected to be conceptually replicated among other groups, but this might be moderated by how stereotypical different groups are and potentially by how favorably they are viewed. Along with marginalized demographics, differing world views and belief systems (such as religious persuasion or political ideology) are often stereotyped, with individuals holding strong beliefs for and against these groups. This would likely moderate how stereotype-consistent judgments of these groups are, but it is again unknown what impact this would have on how accurately members of these groups are judged.

This study was also conducted in a lab, over Zoom, and/or with individuals sitting at a computer answering questions on a screen. These may not fully approximate how most people interact with or form judgments of others, and so future research should find ways to create a more ecologically valid methodology. For example, research could use a round-robin design where participants are brought into the lab and interact with others in-person who they then rate. Previous research has found that there are no real discernable differences in accuracy when judges meet with targets in person or watch a video interaction (Rogers & Biesanz, 2018), but it would be important to see if this is replicated in the context of stereotypes, especially when involving groups that are typically marginalized (e.g., LGBTQ) or that judges hold strong opinions of (e.g., Republican vs Democrat). Judges who hold these strong beliefs may act differently around these targets and therefore elicit different cues for the judge to detect and utilize compared to simply viewing a video interaction.

This study also used individuals' perceptions of how they view groups on average (i.e., individuals were asked to give their perceptions of the average male or female on each item of the BFI-2) and these perceptions may or may not accurately reflect the stereotypes people hold. Research on stereotype accuracy has typically found that when stereotypes are assessed using individuals' perception of the stereotype instead of some more concrete and objective criterion, stereotype accuracy is weaker (Jussim et al., 2018). These perceptions formed the basis for this study and all target personality profiles, normative profiles, and judge ratings were compared against them to see how stereotypical judgments were. Based on work in stereotype accuracy, if this study was conducted using a more objective criterion, stereotypes would have greater accuracy and would probably then have greater utility and more direct relevance to targets. This is consistent with other stereotype research (Jussim et al., 2021), but it is always beneficial to replicate these studies with a more objective criterion where possible.

Finally, one of the main suppositions of this study is that stereotypes and normativity play an intricate part in perceptions of the personalities of others. This was investigated by examining judges' perceptions of targets, but stereotypes and normative judgments may also have impacted the targets' self-ratings and acquaintances' ratings of targets. Normativity was subtracted from the accuracy criterion (i.e., the combined target self-ratings and acquaintance other-reports) before the SAM analysis, which made it possible to study the impact of normativity on these ratings separately from distinctive accuracy. Individuals typically rate themselves and others in a socially desirable way or in a way that is in accordance with perceived norms, and this is accounted for by the normative profile. It also makes sense that because stereotypes have been shown to have high levels of accuracy, stereotype information may have influenced targets- and acquaintance-ratings as well, and thereby impacted the accuracy criterion. Target stereotype consistency was calculated by correlating the accuracy criterion with the stereotype profile. These results show that there was a large range and variability in scores (correlations ranged from -.06 to .79 with a standard deviation of .19). If stereotypes do have a large impact on the accuracy criterion, there should be a large restriction in the range of T-stereotype consistency scores, which is not what was found here. It is possible that for targets who are generally high in T-stereotype consistency, both targets and judges used stereotype information significantly in their ratings, but it would be difficult to tell if this was because targets are stereotype-consistent or if they (and their acquaintances) tend to rate them in a highly stereotypical way. There is some evidence to suggest that stereotypes likely did not have a large impact on self- and acquaintance-judgments because research has found that individuals preferentially encode (are more likely to remember later) stereotype-inconsistent information (compared to stereotype-consistent information) when dealing with in-group members (Koomen & Dijker, 1997), and almost all targets and acquaintances were from a similar demographic group. Future research should look at ways to further pull this apart and directly test how stereotypically targets, acquaintances, and judges are in each of their BFI-2 ratings.

Conclusion

Overall, results for this study were mixed and more research is needed in order to increase the generalizability of claims. There is evidence to suggest that (when judging White non-Hispanic cisgender individuals) accuracy will be typically higher when making judgments that are less stereotype-consistent. It is unclear what causes this because it may be that individuals who are good judges of personality do not have as much of a need to rely on stereotypes because they are sensitive to a target's cues, and so they are both more accurate and less stereotype-consistent. Results indicated that relying on stereotypes overall leads to less accuracy, but if a judge is highly stereotype-consistent, stereotypes may be helpful for accuracy when a target is also highly stereotypical, and a judge makes highly stereotypical judgments. How favorably a judge viewed targets of a specific gender did not seem to impact how accurately targets were rated or how much variability judges had in their ratings. This may be because of the target sample that was used (all White individuals), and it is unclear if this will still be the case for non-White samples.

Important note about bias in judgments

As a final note, the real-life consequence of using stereotypes to make judgments in our daily life need to be considered. This research further supports the idea that, where possible, individuating information should be used, but does not say that stereotype information should be ignored. It is easy to make claims about stereotypes being good or bad, but as with many things in life, this is a nuanced issue where being flexible and sensitive to others is important. Stereotypes should not be completely ignored, but neither should they be the sole piece of information relied on when making judgments. Using stereotypes can have realistic consequences for who individuals will date, trust, hire, arrest, and for other important life outcomes.

Overall, results suggest that judges with less stereotype consistent scores will judge personality with greater accuracy. Whether this is because using stereotypes leads to less accuracy or being a good judge leads to less stereotype use is still an open question. Results do stress the importance of not only relying on stereotypes and instead being sensitive to the individuating information from targets.

References

Adams, H. F. (1927). The good judge of personality. *The Journal of Abnormal and Social Psychology*, *22*(2), 172–181. https://doi.org/10.1037/h0075237

Albers, C. J. (2015). Dutch research funding, gender bias, and Simpson's paradox. *Proceedings of the National Academy of Sciences*, *112*(50), E6828–E6829. https://doi.org/10.1073/pnas.1518936112

Allen, B. P. (1995). Gender stereotypes are not accurate: A replication of Martin (1987) using diagnostic vs. Self-report and behavioral criteria. *Sex Roles*, *32*(9), 583–600. (*note*: see Jussim et al, 2016 which shows that this article actually found the opposite of what the title suggests. Removing one outlier led to a correlation of .61 between sex stereotypes and criteria)

Allport, G. W. (1937). *Personality: A psychological interpretation.*

Allport, G. W., Clark, K., & Pettigrew, T. (1954). *The nature of prejudice*.

American Psychological Association. (1991). In the Supreme Court of the United States: Price Waterhouse v. Ann B. Hopkins: Amicus curiae brief for the American Psychological Association. *American Psychologist*, *46*(10), 1061–1070. https://doi.org/10.1037/0003-066X.46.10.1061

Anderson, N. H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology*, *2*(1), 1. https://doi.org/10.1037/h0021966

Anderson, N. H., & Barrios, A. A. (1961). Primacy effects in personality impression formation. *The Journal of Abnormal and Social Psychology*, *63*(2), 346. https://doi.org/10.1037/h0046719

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290. https://doi.org/10.1037/h0055756

Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. *Cognitive Processes in Stereotyping and Intergroup Behavior*, *1*, 35.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, *54*(7), 462–479. https://doi.org/10.1037/0003-066X.54.7.462

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Beer, A. (2021). Information as a moderator of accuracy in personality judgment. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment*. Oxford University Press.

Berkowitz, L. (1960). The judgmental process in personality functioning. *Psychological Review*, *67*(2), 130–142. https://doi.org/10.1037/h0048565

BetterHelp Editorial Team. (2022). Stereotypes: Definition and why they are wrong | *BetterHelp*. https://www.betterhelp.com

Beyer, S. (1999). The accuracy of academic gender stereotypes. *Sex Roles*, *40*(9), 787–813.

Biesanz, J. C., & Human, L. J. (2010). The cost of forming more accurate impressions: Accuracy-motivated perceivers see the personality of others more distinctively but less normatively than perceivers without an explicit goal. *Psychological Science*, *21*(4), 589–594. https://doi.org/10.1177/0956797610364121

Biesanz, J. C., Neuberg, S. L., Smith, D. M., Asher, T., & Judice, T. N. (2001). When accuracy-motivated perceivers fail: Limited attentional resources and the reemerging self-fulfilling prophecy. *Personality and Social Psychology Bulletin*, *27*(5), 621–629. https://doi.org/10.1177/0146167201275010

Biesanz, J. C., & West, S. G. (2000). Personality coherence: Moderating self–other profile agreement and profile consensus. *Journal of Personality and Social Psychology*, *79*(3), 425–437. https://doi.org/10.1037/0022-3514.79.3.425

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology*, *34*(2), 164–181. https://doi.org/10.1006/jesp.1997.1347

Bollich-Ziegler, K. L. (2021). Self-other knowledge asymmetry (SOKA) model. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment*. Oxford University Press.

Briton, N. J., & Hall, J. A. (1995). Beliefs about female and male nonverbal communication. *Sex Roles*, *32*(1), 79–90.

Brown, J. A., & Bernieri, F. (2017). Trait perception accuracy and acquaintance within groups: Tracking accuracy development. *Personality and Social Psychology Bulletin*, *43*(5), 716–728. https://doi.org/10.1177/0146167217695557

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press.

Campbell, D. T. (1967). Stereotypes and the perception of group differences. *American Psychologist*, *22*(10), 817–829. https://doi.org/10.1037/h0025079

Campbell, I. (2007). Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine*, *26*(19), 3661–3675. https://doi.org/10.1002/sim.2832

Cejka, M. A., & Eagly, A. H. (1999). Gender-stereotypic images of occupations correspond to the sex segregation of employment. *Personality and Social Psychology Bulletin*, *25*(4), 413–423.

Cervone, D., & Shoda, Y. (1999). *The coherence of personality: Social-cognitive bases of consistency, variability, and organization*. Guilford Press.

Chan, M., Rogers, K. H., Parisotto, K. L., & Biesanz, J. C. (2011). Forming first impressions: The role of gender and normative accuracy in personality perception. *Journal of Research in Personality*, *45*(1), 117–120. https://doi.org/10.1016/j.jrp.2010.11.001

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101–129. https://doi.org/10.2307/3001666

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Colman, D. E. (2021). Characteristics of the judge that are related to accuracy. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment*. Oxford University Press.

Colman, D. E., Gibson, J., & Letzring, T. D. (2018). *Motivated accuracy rebooted: A conceptual replication and extension attempt* [Poster]. Society for Personality and Social Psychology, Atlanta, GA, United States.

Colman, D. E., Letzring, T. D., & Biesanz, J. C. (2017). Seeing and feeling your way to accurate personality judgments: The moderating role of perceiver empathic tendencies. *Social Psychological and Personality Science*, *8*(7), 806–815. https://doi.org/10.1177/1948550617691097

Crawford, J. T., Jussim, L., Madon, S., Cain, T. R., & Stevens, S. T. (2011). The use of stereotypes and individuating information in political person perception. *Personality and Social Psychology Bulletin*, *37*(4), 529–542. https://doi.org/10.1177/0146167211399473

Cronbach, L. J. (1955). Processes affecting scores on “understanding of others” and “assumed similarity.” *Psychological Bulletin*, *52*(3), 177–193. https://doi.org/10.1037/h0044919

Davis, M. H., & Kraus, L. A. (1997). Personality and empathic accuracy. In W. J. Ickes (Ed.), *Empathic Accuracy*. Guilford Press.

Ellemers, N. (2018). Gender Stereotypes. *Annual Review of Psychology*, *69*(1), 275–298. https://doi.org/10.1146/annurev-psych-122216-011719

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. https://doi.org/10.1037//0022-3514.38.6.889

Fiske, S. T. (1989). Examining the role of intent: Toward understanding its role in stereotyping and prejudice. *Unintended Thought*, *253*, 283.

Fiske, S. T., & Durante, F. (2016). Stereotype content across cultures: Variations on a few themes. In M. J. Gelfand, C.-Y. Chiu, & Y.-Y. Hong (Eds.), *Handbook of advances in culture and psychology* (Vol. 6, pp. 209–258). Oxford University Press.

Funder, D. C. (1980). On seeing ourselves as others see us: Self–other agreement and discrepancy in personality ratings. *Journal of Personality*, *48*(4), 473–493. https://doi.org/10.1111/j.1467-6494.1980.tb02380.x

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*(1), 75–90. https://doi.org/10.1037/0033-2909.101.1.75

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295x.102.4.652

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology*, *55*(1), 149–158. https://doi.org/10.1037/0022-3514.55.1.149

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology*, *69*(4), 656–672. https://doi.org/10.1037/0022-3514.69.4.656

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*(9), 404–409. https://doi.org/10.1016/j.tics.2004.07.001

Gibson, J. (2019). *The Impact of Prior Information on Personality Judgment Accuracy* [Thesis]. Idaho State University.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Hall, J. A., & Carter, J. D. (1999). Gender-stereotype accuracy as an individual difference. *Journal of Personality and Social Psychology*, *77*(2), 350.

Hall, J. A., Goh, J. X., Mast, M. S., & Hagedorn, C. (2016). Individual differences in accurately judging personality from text. *Journal of Personality*, *84*(4), 433–445. https://doi.org/10.1111/jopy.12170

Halpern, D. F., Straight, C. A., & Stephenson, C. L. (2011). Beliefs about cognitive gender differences: Accurate for direction, underestimated for size. *Sex Roles*, *64*(5), 336–347.

Harris, J. A., Vernon, P. A., & Jang, K. L. (1998). Intelligence and personality characteristics associated with accuracy in rating a co-twin’s personality. *Personality and Individual Differences*, *26*(1), 85–97. https://doi.org/10.1016/S0191-8869(98)00133-0

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*(4), 555–588. https://doi.org/10.1037/a0015701

Haslam, N., Rothschild, L., & Ernst, D. (2002). Are essentialist beliefs associated with prejudice? *British Journal of Social Psychology*, *41*(1), 87–100. https://doi.org/10.1348/014466602165072

Hirschmüller, S., Egloff, B., Schmukle, S. C., Nestler, S., & Back, M. D. (2015). Accurate judgments of neuroticism at zero acquaintance: A question of relevance. *Journal of Personality*, *83*(2), 221–228. https://doi.org/10.1111/jopy.12097

Human, L. J., Biesanz, J. C., Finseth, S. M., Pierce, B., & Le, M. (2014). To thine own self be true: Psychological adjustment promotes judgeability via personality–behavior congruence. *Journal of Personality and Social Psychology*, *106*(2), 286–303. https://doi.org/10.1037/a0034860

Human, L. J., Biesanz, J. C., Parisotto, K. L., & Dunn, E. W. (2012). Your best self helps reveal your true self: Positive self-presentation leads to more accurate personality impressions. *Social Psychological and Personality Science*, *3*(1), 23–30. https://doi.org/10.1177/1948550611407689

Human, L. J., Mignault, M.-C., Biesanz, J. C., & Rogers, K. H. (2019). Why are well-adjusted people seen more accurately? The role of personality-behavior congruence in naturalistic social settings. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000193

Human, L. J., Sandstrom, G. M., Biesanz, J. C., & Dunn, E. W. (2013). Accurate first impressions leave a lasting impression: The long-term effects of distinctive self-other agreement on relationship development. *Social Psychological and Personality Science*, *4*(4), 395–402. https://doi.org/10.1177/1948550612463735

Hyde, J. S. (2014). Gender Similarities and Differences. *Annual Review of Psychology*, *65*(1), 373–398. https://doi.org/10.1146/annurev-psych-010213-115057

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255–259. https://doi.org/10.1016/0165-1765(80)90024-5

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of Personality: Theory and Research*, *3*(2), 114–158.

Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. OUP USA.

Jussim, L. (2018). “Gender Stereotypes are Inaccurate” if You Ignore the Data | Psychology Today. *Psychology Today*. https://www.psychologytoday.com/us/blog/rabble-rouser/201806/gender-stereotypes-are-inaccurate-if-you-ignore-the-data

Jussim, L., Crawford, J. T., Anglin, S. M., Chambers, J. R., Stevens, S. T., Cohen, F., & Nelson, T. D. (2016). Stereotype accuracy: One of the largest and most replicable effects in all of social psychology. In *Handbook of prejudice, stereotyping, and discrimination* (2nd ed., Vol. 2, pp. 31–63).

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (in) accuracy in perceptions of groups and individuals. *Current Directions in Psychological Science*, *24*(6), 490–497. https://doi.org/10.1177/0963721415605257

Jussim, L., Harber, K. D., Crawford, J. T., Cain, T. R., & Cohen, F. (2005). Social reality makes the social mind: Self-fulfilling prophecy, stereotypes, bias, and accuracy. *Interaction Studies*, *6*(1), 85–102.

Jussim, L., Stevens, S. T., & Honeycutt, N. (2018). Unasked questions about stereotype accuracy. *Archives of Scientific Psychology*, *6*(1), 214–229. https://doi.org/10.1037/arc0000055

Jussim, L., Stevens, S. T., & Honeycutt, N. (2021). The Accuracy of Stereotypes About Personality. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment*. Oxford University Press.

Kenny, D. A. (1994). *Interpersonal Perception: A Social Relations Analysis*. Guilford Press.

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*, *102*(3), 390. https://doi.org/10.1037/0033-2909.102.3.390

Kenny, D. A., & La Voie, L. (1984). The Social Relations Model. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 18, pp. 141–182). Academic Press. https://doi.org/10.1016/S0065-2601(08)60144-6

Kenny, D. A., & West, T. V. (2010). Similarity and Agreement in Self- and Other Perception: A Meta-Analysis. *Personality & Social Psychology Review (Sage Publications Inc.)*, *14*(2), 196. https://doi.org/10.1177/1088868309353414

Kenny, D. A., West, T. V., Malloy, T. E., & Albright, L. (2006). Componential Analysis of Interpersonal Perception Data. *Personality and Social Psychology Review*, *10*(4), 282–294. https://doi.org/10.1207/s15327957pspr1004_1

Koomen, W., & Dijker, A. J. (1997). Ingroup and outgroup stereotypes and selective processing. European Journal of Social Psychology, 27(5), 589–601. https://doi.org/10.1002/(SICI)1099-0992(199709/10)27:5<589::AID-EJSP840>3.0.CO;2-Y

Kreft, I. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Sage.

Krzyzaniak, S. L. (2020). *The Role of Target Age in Personality Judgment Accuracy*. Idaho State University.

Krzyzaniak, S. L., & Letzring, T. D. (2021). Characteristics of Traits that are Related to Accuracy of Personality Judgments. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment*. Oxford University Press.

LaPiere, R. T. (1936). Type-Rationalizations of Group Antipathy. *Social Forces*, *15*(2), 232–254. JSTOR. https://doi.org/10.2307/2570963

Latif, N., Human, L. J., Capozzi, F., & Ristic, J. (2022). Intrapersonal Behavioral Coordination and Expressive Accuracy During First Impressions. *Social Psychological and Personality Science*, *13*(1), 150–159.

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality*, *42*(4), 914–932. https://doi.org/10.1016/j.jrp.2007.12.003

Letzring, T. D., Colman, D. E., & Vineyard, J. (2016). *Idaho Test of Accurate Person Perception: Initial Creation of a Standardized Test* [Poster]. Rocky Mountain Psychological Association, Denver.

Letzring, T. D., & Funder, D. C. (2018). Person perception and interpersonal accuracy. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE Handbook of Personality and Individual Differences (Volume 3: Applications of Personality and Individual Differences)* (1 edition). SAGE Publications Ltd.

Letzring, T. D., & Funder, D. C. (2021). The realistic accuracy model. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment*. Oxford University Press.

Letzring, T. D., & Human, L. J. (2014). An examination of information quality as a moderator of accurate personality judgment. *Journal of Personality*, *82*(5), 440–451. https://doi.org/10.1111/jopy.12075

Lewandowski Jr, G., Nardone, N., & Raines, A. (2010). The Role of Self-concept Clarity in Relationship Quality. *Self and Identity*, *9*, 416–433. https://doi.org/10.1080/15298860903332191

Löckenhoff, C. E., Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., De Bolle, M., Costa, P. T., Sutin, A. R., Realo, A., Allik, J., Nakazato, K., Shimonaka, Y., Hřebíčková, M., Graf, S., Yik, M., Ficková, E., Brunner-Sciarra, M., Leibovich de Figueora, N., Schmidt, V., … Terracciano, A. (2014). Gender Stereotypes of Personality: Universal and Accurate? *Journal of Cross-Cultural Psychology*, *45*(5), 675–694. https://doi.org/10.1177/0022022113520075

Lorenzo, G. L., Biesanz, J. C., & Human, L. J. (2010). What Is Beautiful Is Good and More Accurately Understood: Physical Attractiveness and Accuracy in First Impressions of Personality. *Psychological Science*, *21*(12), 1777–1782. https://doi.org/10.1177/0956797610388048

Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*(3), 427–440. https://doi.org/10.1016/j.csda.2003.08.006

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Mackie, M. (1973). Arriving at "truth" by definition: The case of stereotype inaccuracy. *Social Problems*, *20*(4), 431–447.

Mandalaywala, T. M., Amodio, D. M., & Rhodes, M. (2018). Essentialism Promotes Racial Prejudice by Increasing Endorsement of Social Hierarchies. *Social Psychological and Personality Science*, *9*(4), 461–469. https://doi.org/10.1177/1948550617707020

Martin, C. L. (1987). A ratio measure of sex stereotyping. *Journal of Personality and Social Psychology*, *52*(3), 489.

McCauley, C., Stitt, C. L., & Segal, M. (19800101). Stereotyping: From prejudice to prediction. *Psychological Bulletin*, *87*(1), 195. https://doi.org/10.1037/0033-2909.87.1.195

McCauley, C., & Thangavelu, K. (1991). Individual differences in sex stereotyping of occupations and personality traits. *Social Psychology Quarterly*, 267–279.

McCauley, C., Thangavelu, K., & Rozin, P. (1988). Sex stereotyping of occupations in relation to television representations and census facts. *Basic and Applied Social Psychology*, *9*(3), 197–212.

Mignault, M.-C., & Human, L. J. (2021). The good target of personality judgments. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford Handbook of Accurate Personality Judgment*. Oxford University Press.

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality Judgments Based on Physical Appearance. *Personality and Social Psychology Bulletin*, *35*(12), 1661–1671. https://doi.org/10.1177/0146167209346309

Operario, D., & Fiske, S. T. (2004). Stereotypes: Content, Structures, Processes, and Context. In *Social cognition* (pp. 120–141). Blackwell Publishing.

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory, 79*.

Packer, M., & Cole, M. (2015). Culture in Development. In M. Bronstein & M. E. Lamb (Eds.), *Developmental science: An advanced textbook* (7th ed., pp. 43–111). Psychology Press.

Paunonen, S. V., & Kam, C. (2014). The accuracy of roommate ratings of behaviors versus beliefs. *Journal of Research in Personality*, *52*, 55–67. https://doi.org/10.1016/j.jrp.2014.07.006

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*(3), 380–391. https://doi.org/10.1037/0022-3514.61.3.380

Quinn, P. C., Yahr, J., Kuhn, A., Slater, A. M., & Pascalis, O. (2002). Representation of the Gender of Human Faces by Infants: A Preference for Female. *Perception*, *31*(9), 1109–1121. https://doi.org/10.1068/p3331

Ready, R. E., Clark, L. A., Watson, D., & Westerhouse, K. (2000). Self- and Peer-Reported Personality: Agreement, Trait Ratability, and the “Self-Based Heuristic.” *Journal of Research in Personality*, *34*(2), 208–224. https://doi.org/10.1006/jrpe.1999.2280

Reeder, G. D., & Spores, J. M. (1983). *The attribution of morality* (Journal of Personality and Social Psychology, pp. 736–745). https://about.illinoisstate.edu/gdreeder/Documents/Reeder, Spores (1983).pdf

Rogers, K. H., & Biesanz, J. C. (2018). Reassessing the good judge of personality. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000197

Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, *50*(1), 131–142. https://doi.org/10.1037/0022-3514.50.1.131

Ryan, C. (2003). Stereotype accuracy. *European Review of Social Psychology*, *13*(1), 75–109. https://doi.org/10.1080/10463280240000037

Scheier, M. F., Buss, A. H., & Buss, D. M. (1978). Self-consciousness, self-report of aggressiveness, and aggression. *Journal of Research in Personality*, *12*(2), 133–140. https://doi.org/10.1016/0092-6566(78)90089-2

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*(1), 3. https://doi.org/10.1037/0033-295X.99.1.3

Stangor, C. (1995). *Content and application inaccuracy in social stereotyping.* https://doi.org/10.1037/10495-011

Stangor, C. (2016). *The study of stereotyping, prejudice, and discrimination within social psychology: A quick history of theory and research.*

Swim, J. K. (1994). Perceived versus meta-analytic effect sizes: An assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Psychology*, *66*(1), 21.

Taft, R. (1955). The ability to judge people. *Psychological Bulletin*, *52*(1), 1–23. https://doi.org/10.1037/h0044999

Uleman, J. S., & Kressel, L. M. (2013). A brief history of theory and research on impression formation. In D. E. Carlston (Ed.), *Oxford handbook of social cognition* (pp. 53–73).

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403. https://doi.org/10.1037/0033-2909.134.3.383

van der Lee, R., & Ellemers, N. (2015). Gender contributes to personal research funding success in The Netherlands. *Proceedings of the National Academy of Sciences, 112*(40), 12349–12353. https://doi.org/10.1073/pnas.1510159112

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300. https://doi.org/10.1037/a0017908

Verkuyten, M. (2003). Discourses about ethnic group (de-)essentialism: Oppressive and progressive aspects. *British Journal of Social Psychology*, *42*(3), 371–391. https://doi.org/10.1348/014466603322438215

Waggoner, A. S., Smith, E. R., & Collins, E. C. (2009). Person perception by active versus passive perceivers. *Journal of Experimental Social Psychology*, *45*(4), 1028–1031. https://doi.org/10.1016/j.jesp.2009.04.026

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–263. https://doi.org/10.1016/j.tics.2009.03.006

Ybarra, O. (2001). When first impressions don't last: The role of isolation and adaptation processes in the revision of evaluative impressions. *Social Cognition*, *19*(5), 491–520. https://doi.org/10.1521/soco.19.5.491.19910

Ybarra, O., Schaberg, L., & Keiper, S. (1999). *Favorable and unfavorable target expectancies and social information processing* (Vol. 77). https://doi.org/10.1037/0022-3514.77.4.698

Zaki, J., & Ochsner, K. (2011). Reintegrating the Study of Accuracy Into Social Cognition Research. *Psychological Inquiry*, *22*(3), 159–182. https://doi.org/10.1080/1047840X.2011.551743

Appendix A (Measures)

The Big Five Inventory–2 (BFI-2)

Self-report instructions: Here are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who *likes to spend time with others?* Please select a number next to each statement to indicate the extent to which you agree or disagree with that statement.

Stereotype criterion instructions: Here are a number of characteristics that may or may not apply to the groups you are rating. For example, do you agree that average individuals in this group are people who *like to spend time with others?* Please select a number next to each statement to indicate the extent to which you agree or disagree with that statement.

Acquaintance instructions: Here are a number of characteristics that may or may not apply to your acquaintance. For example, do you agree that they are someone who *likes to spend time with others?* Please select a number next to each statement to indicate the extent to which you agree or disagree with that statement.

Judge other report instructions: Here are a number of characteristics that may or may not apply to the person you just watched in the video. For example, do you agree that they are someone who *likes to spend time with others?* Please select a number next to each statement to indicate the extent to which you agree or disagree with that statement.

1	**2**	**3**	**4**	**5**
Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree Strongly

[***I am, The person is, I see*** **average individuals in this group as,** ***someone who...***

1. Is outgoing, sociable.
2. Is compassionate, has a soft heart.
3. Tends to be disorganized.
4. Is relaxed, handles stress well.
5. Has few artistic interests.
6. Has an assertive personality.
7. Is respectful, treats others with respect.
8. Tends to be lazy.
9. Stays optimistic after experiencing a setback.
10. Is curious about many different things.
11. Rarely feels excited or eager.
12. Tends to find fault with others.
13. Is dependable, steady.
14. Is moody, has up and down mood swings.
15. Is inventive, finds clever ways to do things.
16. Tends to be quiet.
17. Feels little sympathy for others.
18. Is systematic, likes to keep things in order.
19. Can be tense.
20. Is fascinated by art, music, or literature.
21. Is dominant, acts as a leader.
22. Starts arguments with others.
23. Has difficulty getting started on tasks.
24. Feels secure, comfortable with self.
25. Avoids intellectual, philosophical discussions.
26. Is less active than other people.
27. Has a forgiving nature.
28. Can be somewhat careless.
29. Is emotionally stable, not easily upset.
30. Has little creativity.
31. Is sometimes shy, introverted.
32. Is helpful and unselfish with others.
33. Keeps things neat and tidy.
34. Worries a lot.
35. Values art and beauty.
36. Finds it hard to influence people.
37. Is sometimes rude to others.
38. Is efficient, gets things done.
39. Often feels sad.
40. Is complex, a deep thinker.
41. Is full of energy.
42. Is suspicious of others' intentions.
43. Is reliable, can always be counted on.
44. Keeps their emotions under control.
45. Has difficulty imagining things.
46. Is talkative.
47. Can be cold and uncaring.
48. Leaves a mess, doesn't clean up.
49. Rarely feels anxious or afraid.
50. Thinks poetry and plays are boring.
51. Prefers to have others take charge.
52. Is polite, courteous to others.
53. Is persistent, works until the task is finished.
54. Tends to feel depressed, blue.
55. Has little interest in abstract ideas.
56. Shows a lot of enthusiasm.
57. Assumes the best about people.
58. Sometimes behaves irresponsibly.
59. Is temperamental, gets emotional easily.
60. Is original, comes up with new idea

Item numbers for the BFI-2 domain scales are listed below. Reverse-keyed items are denoted by "R." For more information about the BFI-2, visit the Colby Personality Lab website (http://www.colby.edu/psych/personality-lab/).

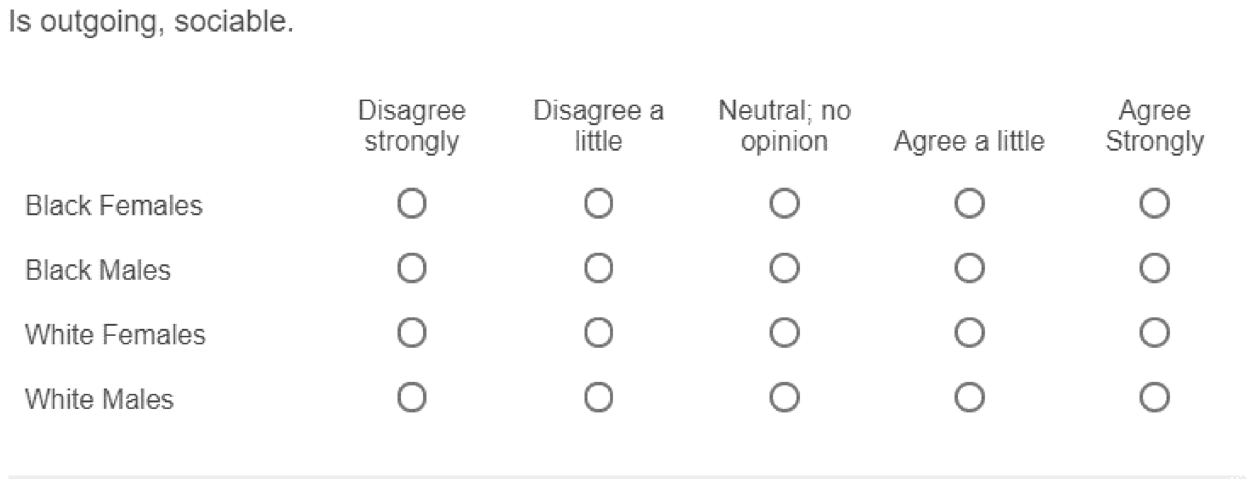
Domain Scales

Extraversion: 1, 6, 11R, 16R, 21, 26R, 31R, 36R, 41, 46, 51R, 56
Agreeableness: 2, 7, 12R, 17R, 22R, 27, 32, 37R, 42R, 47R, 52, 57
Conscientiousness: 3R, 8R, 13, 18, 23R, 28R, 33, 38, 43, 48R, 53, 58R
Negative Emotionality: 4R, 9R, 14, 19, 24R, 29R, 34, 39, 44R, 49R, 54, 59
Open-Mindedness: 5R, 10, 15, 20, 25R, 30R, 35, 40, 45R, 50R, 55R, 60

Example BFI-2 questions for stereotype profile study

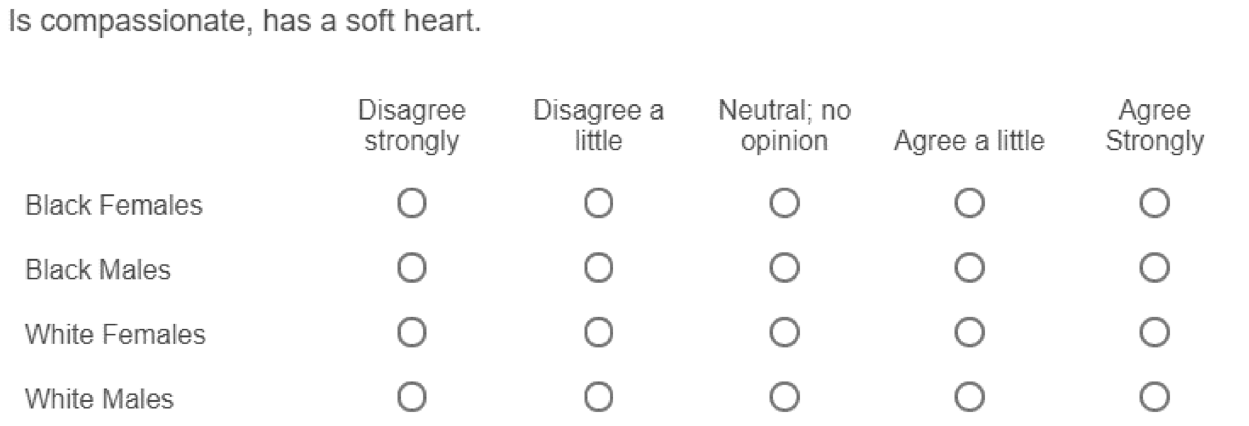
Below you will see a list of characteristics that may or may not apply to each group. For example, do you agree that on average members of this group are individuals who like to spend time with others? Please indicate the extent to which you agree or disagree with each statement regarding each group.

I see the average individual in this group as someone who …



Is outgoing, sociable.

	Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree Strongly
Black Females	○	○	○	○	○
Black Males	○	○	○	○	○
White Females	○	○	○	○	○
White Males	○	○	○	○	○



Is compassionate, has a soft heart.

	Disagree strongly	Disagree a little	Neutral; no opinion	Agree a little	Agree Strongly
Black Females	○	○	○	○	○
Black Males	○	○	○	○	○
White Females	○	○	○	○	○
White Males	○	○	○	○	○

Appendix B (Email Templets)

Stereotype Profile Creation

Thank you for agreeing to take part in this study. The purpose of this study is to better understand how personality perceptions are made. We expect this study will take about 30 minutes to complete. Before clicking on the link below, please remove any and all possible distractions. This includes moving or turning off all electronic devices and notifying any individuals who may distract you that you need to be uninterrupted for the next 30 minutes. If it is likely that you will be disrupted over the next 30 minutes, we ask that you please postpone taking this study until you can commit 30 uninterrupted minutes to this study. Thank you again for your participation. Please click the link below when you are ready. You will first see a short form that will notify you of the purpose of this study and your rights as a participant, after which you will begin the study. If you have any questions before, during, or after your participation in this study, please contact Jacob Gibson at gibsjaco@isu.edu or at (719) 440-1470.

[LINK TO THE STUDY]

Stimuli Creation (Acquaintances)

[name of acquaintance] I have just taken part in a psychology study about personality perceptions. As part of this study, they are asking me to contact individuals close to me who I feel could accurately rate my personality. You will find information about the study below.

Thank you for agreeing to take part in this study. The purpose of this study is to help us better understand how personality perceptions are made. Your participation is important for the success of this study so please participate if you are at all able. We expect this study will take about 30 minutes to complete. Before clicking on the link below, please remove any and all possible distractions. This includes moving or turning off all electronic devices and notifying any individuals who may distract you that you need to be uninterrupted for the next 30 minutes. Thank you again for your participation. Please click the link below when you are ready. You will first see a short form that will notify you of the purpose of this study and your rights as a participant, after which you will begin the study. If you have any questions before, during, or after your participation in this study, please contact Jacob Gibson at gibsjaco@isu.edu or at (719) 440-1470.

[LINK TO THE STUDY]

Judgments

Thank you for agreeing to take part in this study. The purpose of this study is to help us better understand how personality perceptions are made. We expect this study will take about 1 hour and 30 minutes to complete. You will receive an email 2-3 days prior to your assigned time slot that will contain a link to the zoom meeting you will attend. This email will also contain information about how to set up and use zoom if you have not done so before.

Before your assigned time, please remove any and all possible distractions. This includes moving or turning off all electronic devices that will not be used during the study and notifying any individuals who may distract you that you need to be uninterrupted for your hour and a half time slot. If it is likely that you will be disrupted during your assigned time slot, please reschedule for a time when you are less likely to be distracted.

As part of this study, you will be watching the videos of up to six previously recorded interviews. Please make sure there is no one in earshot of these videos except for you, or please have a pair of headphones ready for this study.

Thank you again for your participation. If you have any questions before or after your participation in this study, please contact Jacob Gibson at gibsjaco@isu.edu or at (719) 440-1470.

Appendix C (Video Transcript)

Introduction video transcript

For this experiment, you will watch 6 videos of individuals interacting with another person who is not on camera. Before watching any videos, you will be asked a few questions about your own personality. After each video, you will answer questions about that person. Videos are approximately 6 to 7 minutes in length. After watching all the videos, you will answer several questions about yourself. During the experiment, a research assistant will be available to answer questions about the procedures but not about the content of the videos. When the videos are playing please do your best to pay attention in order to properly answer the questions that follow. Before you watch any of the videos, you will see pictures of all 6 individuals. If you recognize or know the people in any of the images, please indicate it on the survey, and you will be assigned to a different set of individuals.

We would like to inform you that we check responses carefully in order to make sure that people read the instructions for the task and respond carefully. We can only use data from participants who clearly demonstrate that they have read and understood the questionnaires and tasks. There will be some very simple questions throughout the experiment that test whether you are reading the instructions and responding carefully. Please be sure to answer these correctly.

Please remove any and all possible distractions. This includes moving or turning off all electronic devices that will not be used during the study and notifying any individuals who may distract you that you need to be uninterrupted for your hour and a half time slot.

If you have any questions during the study, please use the chat feature in zoom to notify the research assistant. Thank you again for your participation.

Appendix D (Target Interview Questions)

1. Tell me a little bit about yourself. How would you describe yourself to someone who has never met you?
2. What are some of your hobbies or activities you enjoy doing?
    a. Why do you think you enjoy these types of activities?
    b. What hobbies do you see yourself pursuing in the future?
3. What is a goal you are currently working toward?
    a. What success have you had with this goal so far?
    b. What challenges have you had with this goal so far?
4. What is an example of an especially meaningful moment in your life?
    a. What made this moment meaningful to you?
    b. How did this moment impact your life?
5. What is a challenge or conflict you are currently facing?
    a. How do you feel about it?
    b. What are you doing to deal with it?

---------------------- End of part 1 of the interview, the following questions will be asked but will not be a part of this dissertation project --

6. When you first meet someone, what information do you think you use to help you make judgments about their personality?
7. What is your race and gender?
    a. What are some common personality stereotypes about [race and gender] that you are familiar with?
    b. How well do you think these personality stereotypes describe you?
        i. What do they get right? What do they get wrong?
    c. How well do you think these personality stereotypes describe someone who is a [race and gender]?
        i. What do they get right? What do they get wrong?

8. To what extent do you think people use personality stereotypes when making judgments of another person's personality?
9. Do you think you ever use personality stereotypes when making judgments of another person's personality? If so, in what ways?
10. When would it be useful to use stereotypes to make judgments of others?
11. When would it not be useful to use stereotypes to make judgments of others?