Does it Reinforce or Ridicule?

Predicting Inferences of Endorsement vs. Subversion in Stereotype Humor

by

Erika B. Langley

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2024 by the
Graduate Supervisory Committee:

Michelle "Lani" Shiota, Chair
Steven Neuberg
Gene Brewer
Keith Maddox

ARIZONA STATE UNIVERSITY

May 2024

ABSTRACT

Stereotype humor is a highly prevalent but particularly divisive phenomenon, with the potential for both negative and positive social implications. While highly subjective, interpretations of stereotype humor's subtext (support/challenge of stereotype) have major implications for reactions to this type of humor. This experimental study (N = 104) represents a novel investigation of the effect of two facets of stereotype humor, explicitness of stereotyping and stereotype distortion, on judgments of stereotype endorsement (support) versus subversion (challenge) in memes about four different groups (Asian, Hispanic, Irish, White) and associated group stereotypes. In this completely within-subjects design, participants viewed several memes about the target groups which varied systematically by the two factors of interest and provided judgments of stereotype endorsement versus subversion, ratings of funniness, and ratings of offensiveness. Multilevel models were used to determine the effect of explicitness, distortion, and their interaction, as well as target group, on judgements of stereotype humor while accounting for nesting of responses within participants. Results showed that stereotype distortion (e.g., exaggeration) and explicit stereotyping (e.g., overtly linking group to stereotype) both significantly predicted greater ratings of subversion. Unexpectedly, stereotype distortion also predicted greater levels of offense. Interestingly, marginalized group membership (i.e., Asian, Hispanic) significantly predicted lower ratings of subversion, lower funny ratings, and higher offense ratings. Findings highlight the significant role of explicitness and distortion when considering how individuals interpret the subtext of stereotype humor. Furthermore, findings underscore the major influence of group status on judgments and social implications of this type of humor.

Overall, this study contributes to a better understanding of the mechanisms by which individuals interpret stereotype humor, providing valuable insights for promoting better intergroup relations and communication.

DEDICATION

I dedicate this my mother, who inspired my pursuit of higher education in the face of all challenges; Neil, who kept me applying and who I wish was here to see what all of his support helped me to achieve; and to my wonderful husband Matt, the funniest and smartest person I know, who has helped me grow into the academic I am today.

ACKNOWLEDGMENTS

Thank you to Lani for encouraging me to be the best researcher I can be, providing excellent guidance on how to succeed in academia, allowing me the flexibility to study such a cool and complex topic, and for putting up with my weird sense of humor.

Thank you to my committee for believing in my vision and providing invaluable mentorship. I am deeply honored to have been selected as the recipient for the Robert B. Cialdini Dissertation Project Prize in Social Psychology by Bob and the Social Area Faculty for this work. Thank you to the Cognitive Area Faculty for being my home away from home. I have learned so much from all of you and I am profoundly grateful for your support over the years. My future students thank you as well.

Thank you to Alex for being my lifeboat when I was drowning in complicated statistical analyses. Your expertise and reassurance provided a sense of relief I cannot put into words. I look forward to future humor collaborations with you.

Thank you to all the people who encouraged me through this long and arduous process. It truly takes a village.

And finally, in the wise words of Snoop Dogg, 'I want to thank me for believing in me. I want to thank me for doing all this hard work. I want to thank me for never quitting. I want to thank me for being me at all times.'

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER 1

DOES IT REINFORCE OR RIDICULE?

"If anybody's upset about that joke, just tweet about it on your phone that was also made by these kids."

– *Chris Rock*

During the 2016 Oscars, host Chris Rock brought three Asian American kids on stage with him for a joke, introducing them as PricewaterhouseCoopers accountants ("Ming Zhu, Bao Ling, and David Moskowitz") in charge of tabulating the Oscar votes, referencing the stereotype that Asians are hardworking and good at math. Rock followed this with a satirical comment alluding to the stereotype of child labor in Asia, intended as a preemptive challenge to any social media outrage over his jokes. While many in the audience were heard laughing, many at-home viewers were offended, and Rock faced much online criticism and backlash for his stereotype jokes. The stark contrast between audience laughter and online outrage is characteristic of stereotype humor, which often elicits mixed emotional reactions from listeners.

Stereotype humor is highly prevalent but can be particularly divisive (Schoem, 2003; Lockyer & Pickering, 2005; Borgella et al., 2019). It is generally considered taboo for its association with stereotyping and prejudice. Decades of research have established the negative consequences of stereotyping (Lu et al., 2015), the distressing nature of conversations surrounding identity (Richeson & Shelton, 2007), as well as the use of stereotypes to justify discrimination (Dovidio & Gaertner, 2010). More and more in our society, overt expressions of prejudice are on the decline and individuals do not want to be perceived as prejudicial (Sydell & Nelson, 2000; Devine et. al., 2002). Laughing is generally a signal of social acceptance or approval (Provine, 2000), as such, laughing at a

stereotype joke could be interpreted by others as agreement with the stereotype or setting a norm that prejudiced attitudes are acceptable (Ford & Ferguson, 2004).

However, humor can facilitate social bonds with others and there are cases in which joking about stereotypes can have a positive social impact. A stereotype joke can be used to bring two individuals closer together through the communication of shared experience or signaling a sense of similar worldview (Pratt, 1998; Rappaport, 2005; Flamson & Barrett, 2008). Alternatively, self-directed stereotype humor can be leveraged to better navigate challenging interpersonal hierarchical relationships (Martin, 2004). In other cases, individuals might laugh at stereotype jokes because the content is relatable, and it helps to cope with stigma (Hart, 2007). This type of stereotype humor is often used to foster pride in a positive ethnic identity (Leveen, 1996). There is also an opportunity to dismantle stereotypes by using humor to directly challenge them (Kramer, 2020).

Humor is a form of ambiguous communication with multiple possible interpretations. The risks and rewards associated with stereotype humor are strongly linked, in theory, to how individuals interpret the underlying subtext of the joke with regards to the stereotype (Rappaport, 2005; Saucier et al., 2016). In some cases, stereotype humor can be perceived as endorsing or *reinforcing* the validity of stereotypes; endorsement of stereotypes, particularly negative stereotypes, is linked to prejudice and stigmatization of marginalized groups. In other cases, stereotype humor can be interpreted as subversive or *ridiculing* the stereotype. This interpretation is far more prosocial and taken as a defense against prejudice and discrimination. But how do people decide whether a stereotype joke is endorsing or subverting a stereotype?

CHAPTER 2

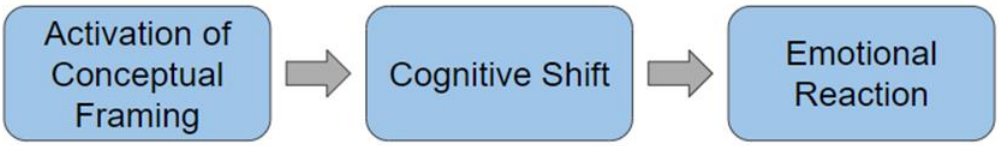
INTRODUCTION

The Humor Process

Humor, in all its varied forms, is a fundamental human experience. To understand how individuals respond to stereotype humor it is important to consider the process from which these responses result. At a rudimentary level, humor is cognitive. Here I define a humor episode as the psychological process by which some initially established conceptual framing of a situation undergoes a cognitive shift in meaning to an alternate, unrelated, yet contextually appropriate interpretation, resulting in amusement. In my model, presented in Figure 1 below, I break this down into three core stages: activation of some initial conceptual framework, cognitive shift, followed by an emotional response.

Figure 1.

The Humor Process.



This model synthesizes several past theories of humor across a wide range of disciplines (Koestler, 1964; Suls, 1972; Raskin, 1985; Attardo, 1997; Veatch, 1998; Beeman, 1999; Latta, 1999; Gervais & Wilson, 2005; Kulka, 2007; Martin & Ford, 2018; Flamson & Barrett, 2008; Dynel, 2009; Kao et al., 2016; Gabora & Kitto, 2017; Raizada, 2020; Warren et al., 2021); the model is designed to be general enough to account for both highly structured (e.g., jokes, performance humor) and more spontaneous (e.g., situational, unintentional) instances of humor.

Activation of a Conceptual Framework.

The first stage of a humor episode requires the activation of a network of concepts, schema, or some conceptual framing suitable for the current situation. This process occurs through spreading activation (Anderson,1983; Collins & Loftus, 1975; McNamara, 2005). According to models of spreading activation, concepts are represented as interconnected nodes in a network structure. When one node (concept, schema) is activated, this activation then spreads to connected, or semantically related, nodes; more related concepts will be more strongly activated. Nodes become activated when we directly perceive the stimulus it represents (looking at a mug will activate the concept "mug") or when the concept is the goal of some cognitive activity (being instructed to remember a sunset will activate the concept of "sunset").

Distributed network models of spreading activation build on this framework by modeling concepts as memory patterns built from collections of distributed representations (rather than single nodes) that once activated further activate correlated conceptual patterns (Plaut & Booth, 2000; Lerner et al., 2012). For example, thinking about having to get up in the morning will also activate thoughts about related concepts like making coffee or having breakfast. These categorized conceptual networks are the schematic representations stored in long-term memory which, when activated, guide cognitive processing, and can aid in interpreting the world around us (Hastie, 1981).

Previous theoretical work on humor often references this initial stage as "the setup" (Attardo, 1997; Beeman, 1999; Dynel, 2009). In the most prototypical example of a humor elicitor, a constructed joke, the setup establishes the subject matter; information is presented to a listener with the intent to build up the conceptual frameworks relevant

for the joke. When one encounters a setup, each component activates associated concepts through spreading activation.

Cognitive Shift.

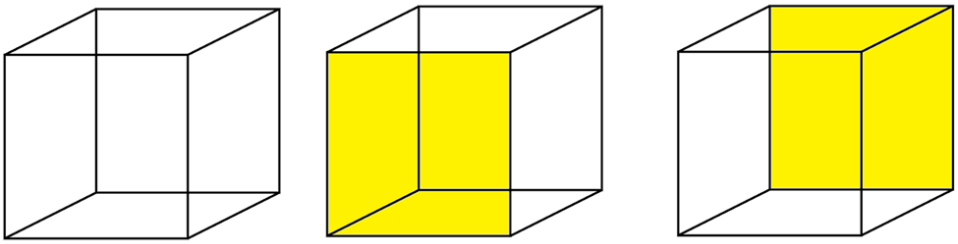
The next stage in the humor process occurs when attention is rapidly and automatically shifted to an alternate conceptual framing that makes logical sense, but which is incompatible with the initial framing. While the setup establishes multiple viable conceptual framings, only one interpretation is considered initially, that which best fits the situation at the time (McBeath, 2018). The cognitive shift is initiated when we observe some aspect or feature of what we are attending to that triggers a reconsideration of the global interpretation we initially felt was most probable for one of the competing alternate interpretations (Raskin, 1985; Latta, 1999). This phenomenon has also previously been referred to as incongruity (Kao et al., 2016), simultaneity (Attardo & Raskin, 1991; Veatch, 1998), bisociation (Koestler, 1964; Gabora & Kitto, 2017) script opposition (Raskin et al., 2009), and juxtaposition (Warren et al., 2021). Broadly, there are contrasting interpretations of the same stimulus held at the same time.

Theorists have conceptualized this cognitive shift as a bifurcation point during which the alternate competing conceptual framing becomes more strongly activated (Gabora & Kitto, 2017). In terms of spreading activation, this means that the initial framing and the one initiated by the cognitive shift (a more distant, loosely related concept) have both reached the threshold for conscious awareness, creating a bimodal distribution of activation rather than a single peak. This bifurcation is akin to perceptual bistability, a phenomenon during which we flip back and forth between multiple interpretations of an unaltered stimulus (Denham et al., 2018). As with a Necker cube,

which allows for multiple interpretations for which side a hollow cube is facing (Figure 2; Necker, 1832), the cognitive shift initiates a point of conceptual bistability.

Figure 2.

Perceptual Bistability of a Necker Cube.



Note. A Necker cube can be viewed as facing left or right. Both interpretations exist simultaneously. Humor relies on a similar bistability of concepts.

Cognitive shifts during humor episodes have previously been characterized phenomenologically as moments of insight or the experience of an "Aha! moment" (Kounios & Beeman, 2009). This is a cognitive process by which unconscious reorganization of information surrounding some problem results in the sudden comprehension of a solution (Kounios & Beeman, 2014). In the case of a humor episode, there is a sudden awareness of the competing conceptual framing. As is the case with insight, when effortful search strategies are required to find the solution to a joke, there is a lack of sudden comprehension, and the humor experience is diminished. Additionally, if one is unable to shift to an alternative framing, the humor process is interrupted.

Emotional Reaction with Social Implications.

The resulting emotional response associated with humor (amusement) is highly social in nature. Researchers examining the daily occurrence of laughter, the social play vocalization often accompanying humor episodes, have noted that most laughter is

elicited by and in the presence of others (Martin & Kuiper, 1999). Humor and laughter occur in all cultures, with laughter universally recognized to communicate both a sense of playfulness and social acceptance or approval (Apte, 1985; Provine, 2000). Humor is frequently used as a tool to foster positive social interactions and positive emotions in others, not only with established long-term social partners like romantic partners (Hall, 2013; Kurtz & Algoe, 2015) and friends (Knight, 2013), but also with strangers (Fraley & Aron, 2004). Similarly, at the group level, humor is theorized to solidify in-groups by functionally reinforcing group norms and boundaries (Hay, 2000; Fine & Soucey, 2005). Thus, humor can be utilized to build upon existing relational foundations through reinforcing a sense of shared reality (Flamson & Barrett, 2008).

Much work has established the capacity for humor to help us bond with others, however, humor can also function to push others away (Martineau, 1972; Hay, 2000). This can range from the level of relatively harmless teases to outright derogation. In the case of teasing, individuals use playful provocation to make someone acutely aware that they are deviating from some social norm or desired state (Keltner et al., 2001). Teasing encompasses a diverse set of behaviors that function to communicate negative information about the target (norm violation) in a less threatening way. Interestingly, there are key differences in what teasers and targets of teases attend to during a teasing exchange. Teasers tend to highlight the affiliative or affectionate nature of their remarks, while targets are more aware of the underlying negative messaging (Shapiro et al., 1991).

Even more aggressive, disparagement humor – attempts to amuse through the denigration of others – is an example of a far more antisocial form of humor (Ford et al., 2015). Whereas teasing incorporates affection, disparagement humor is far more

derogatory in nature (e.g., racist, sexist, homophobic jokes); initiators are able to avoid negative social recourse by falling back on the non-serious, humorous framing (Ford & Ferguson, 2004). What these more negative instantiations of humor highlight is the importance of considering both the literal content of a joke and the underlying message being communicated when investigating reactions to humor.

Stereotype Humor

What happens when the concepts activated in a joke are group stereotypes? Stereotype humor is the umbrella term for any humor which references or evokes some group stereotype. Stereotypes are defined as relatively inflexible, overgeneralized beliefs about a particular group of people (Schneider, 2005; Neuberg et al., 2020). These beliefs encompass distinguishing qualities associated with group members, expected social roles, and coalesce to form a template of a typical group member; these schematic representations of group information are held in long term memory and become automatically activated when we are exposed to relevant group information (Fiske & Neuberg, 1990; Stangor & Lange, 1994).

When one thinks about stereotypes and stereotyping, the first association is typically with race. While humor about racial and ethnic stereotypes has received much attention in the literature (Maio et al., 1997; Weaver, 2010; Saucier et al., 2016), researchers have also examined humor about gender (often termed sexist humor; Ford et al., 2008; Romero-Sanchez et al., 2010; Kochersberger et al., 2014; Weaver et al., 2014), religion (Ford et al., 2014), sexual orientation (O'Connor et al., 2017; Thai et al., 2019), and political affiliation (Braun & Preiser, 2013). Group membership can extend beyond these highly valued social identities to rather arbitrary (yet meaningful) distinctions (e.g.,

Apple vs. Android phone users). I use the umbrella term stereotype humor to avoid the finer-grain distinctions often used previously, which often assume harmful intent (i.e., disparagement humor, sexist humor, racial humor, anti-gay humor).

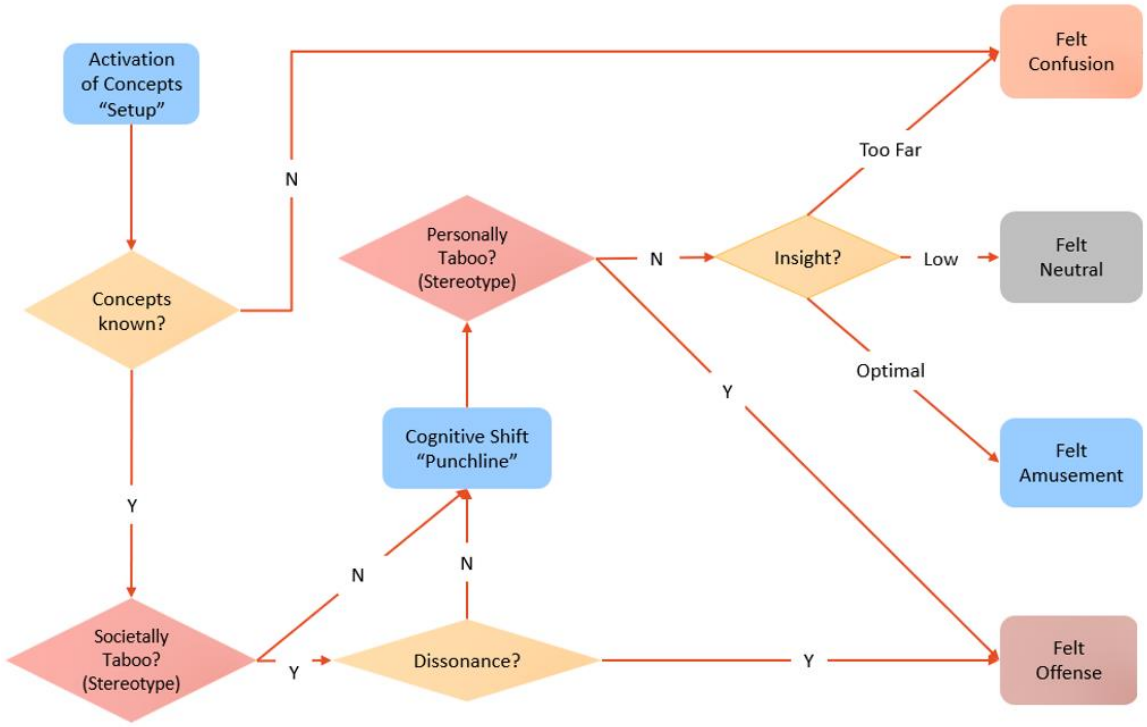
Given the rather ambiguous nature of humor, it is important to establish the factors which influence individuals' reactions to and interpretations of stereotype humor. Figure 3 depicts how stereotypes can enter and alter the basic humor process. This expanded model first considers whether the listener is aware of concepts initiated by the setup. If concepts are unknown, the process ends with confusion. If concepts are known, the next thing to consider is whether the concepts are perceived by the listener as societally or normatively taboo. Since the concepts are being shared by another individual, the listener must consider the normative schematic representation of the concepts in the setup (i.e., "what do people think about this topic?"). While this model mainly considers stereotypes, it should also apply to joking about other taboo topics.

If concepts are not taboo then the process can continue to the cognitive shift (punchline), however, if they are considered culturally taboo then the next decision point is whether thinking about the concept initiates dissonance in the listener. Dissonance in this case is a sense of tension elicited by a topic that is too rigid, too cognitively demanding, or too outside the realm of normal conversation to facilitate a cognitive shift; in other words, topics about which people tend to say "you can't joke about", including stereotypes. When there is a sense of felt dissonance, individuals are likely to feel offended before even getting to the punchline. In some cases, though, the level of dissonance is manageable and the listener is able to get to the punchline.

The punchline is another point in the model at which stereotypes can enter, whether or not the stereotype was mentioned in the setup. At this point, rather than societal taboo, individuals consider how personally taboo the punchline is with regard to the stereotype because the cognitive shift is a personal interpretation of the intent of the joke. If the interpretation is personally concerning or taboo, offense is the likely emotional outcome. If the concept is not offensive, then the emotional reaction is dependent on the level of insight ("aha moment") felt from the cognitive shift. If the shift is too far or too effortful, a listener might feel confused. If the shift is not far enough or too low, a listener will likely recognize that a joke was attempted but feel neutral. Amusement should follow an optimal cognitive shift.

Figure 3.

Expanded Stereotype Humor Process Model



Research has investigated the impact of several social factors on resulting emotional reactions to stereotype humor. The first is *joker identity*. We are generally biased to view someone discussing or criticizing their own group as more acceptable than cases when commentary is coming from someone not of the group (Sutton et al., 2006). This extends to joking; people react more favorably to someone joking about their own group. Across three studies, Thai and colleagues (2019) found that stereotype humor—specifically disparagement humor about sexual orientation, race, and gender—was evaluated more positively (more acceptable, less offensive, funnier) when the joker belonged to the group being disparaged. This in-group joker advantage effect was also replicated by Langley and Shiota (in prep). In this work, joker group status (Asian vs. White) exhibited the greatest influence on perceptions of a joke about Asian parents being demanding; participants rated ingroup (Asian) jokers as funnier, less offensive, and leading to greater closeness between conversation partners than outgroup (White) jokers. Additional work demonstrates that factors surrounding group composition (hierarchy; Keltner et al., 1998; Robinson & Smith-Lovin, 2001; mixed vs. homogenous group setting; Langley & Shiota, in prep) also have unique influence on emotional reactions to stereotype humor. Considering the expanded stereotype humor process model, the joker's identity moderates the level of dissonance individuals feel when listening to a joke about some group stereotype. In-group jokers should lead to lower levels of dissonance, allowing listeners to continue on with the humor process, whereas outgroup jokers increase dissonance and likelihood of offense.

Beyond joker identity, *stereotype valence* also influences reactions to stereotype humor. Stereotype valence refers to the overall tone of the characterization of some group

feature. Valence can be negative ("Women are bad drivers"), positive ("Asians excel academically"), or neutral ("New Zealanders like rugby") depending on the implications being put forth by the stereotype. Xing (2019) replicates the ingroup advantage effect and provides evidence for the differential reception of positive versus negative stereotype jokes; in this work, negative stereotype jokes about Asian-Americans (bad driving) were reported to be more offensive but more amusing; the inverse was true for positive Asian-American stereotype (success driven) jokes. Generally, presenting an indisputably negative stereotype about a group in an attempt to evoke laughter is associated with disparagement and viewed as harmful for implying prejudice (Ford & Ferguson, 2004). Less is known regarding stereotype humor about more ambiguous stereotypes that fall within a slightly negative to rather neutral range. In the expanded stereotype humor model, stereotype valence can influence a listener's perception of the level of societal taboo (i.e., negative stereotypes are more taboo than positive) and also how personally taboo they find the cognitive shift (i.e., doubling down on a negative racial stereotype in the punchline is more taboo than challenging a negative racial stereotype, unless the listener is prejudiced against the group).

Finally, judgments of stereotype humor are also modulated by *social concern*, or our beliefs of how laughing at the joke will impact how others perceive our character. Follow up studies by Langley and Shiota (in prep) investigated the moderating effect of trait concern over appearing prejudiced, as measured by the Concern subscale of the Motivation to Control Prejudiced Reactions scale (MCPR Concern subscale; Dunton & Fazio, 1997) Interestingly, Pages & Shiota found that when personal concern over appearing prejudiced was taken into account, the effect of joker group status was

eliminated for participants that were low in concern. These individuals reported no difference in offense whether the joker was ingroup or outgroup with regard to the stereotyped group targeted in the joke. However, participants higher in concern over appearing prejudiced demonstrated an amplification of the ingroup advantage effect, judging outgroup jokers much harsher (less funny, more offensive) than ingroup jokers.

This personal concern is relevant at two inflection points in the expanded stereotype humor process model. First, concern over appearing prejudiced should positively correlate with the level of dissonance felt when a normatively taboo subject is stated in the setup; when concern (and therefore dissonance) is high, the likely emotional response is offense. If concern is low, the process can continue to the punchline without interruption. At this point, one's level of concern can influence the way a stereotype is interpreted in the punchline. When considering the "sword and shield" metaphor of racial humor (Rappaport, 2005), concern over appearing prejudice may be strongly tied to how individuals interpret the underlying messaging of a stereotype joke (i.e., whether a joke endorses or subverts a stereotype); jokes that are viewed as subverting a stereotype should be less concerning or less socially consequential than jokes that endorse a stereotype. However, the process by which interpretations of endorsement versus subversion are made is rather understudied.

Endorsement versus Subversion

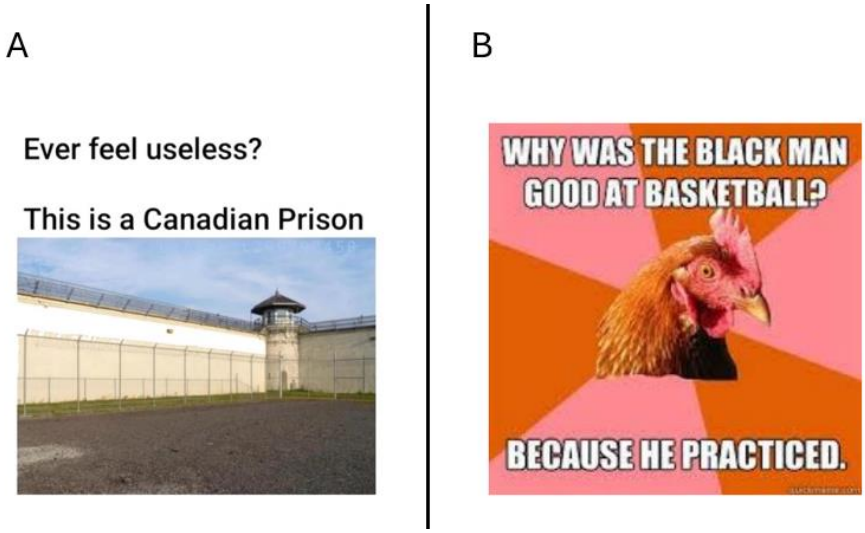
Stereotype humor that *endorses* functions to support the stereotype referenced within the joke or reinforce some existing status differential between groups (Holmes & Marra, 2002). These types of jokes "maintain the status quo" and uphold the stereotype in the listener's mind. Consider the following example in Panel A of Figure 4. Mentioning

Canadians activates associated group stereotypes, primarily that Canadians are exceedingly nice (Snefjella et al., 2018). While not overtly mentioned in the joke, in order to make sense of a Canadian prison being useless and empty, the viewer must be aware of the Canadian stereotype. While exaggerated to quite an extreme (i.e., Canadians are *so* nice their prisons are empty), the humor relies on the listener having to call to mind the stereotype and process the information at face value in order to understand the joke.

Critically, there is no need to believe or personally endorse the stereotype as accurate for the listener to comprehend the joke, they just need to know that the stereotype exists. Previous research has conflated jokes that endorse stereotypes with disparagement (Miller et al., 2019), however, this does not account for situations in which the joke endorsing a negative stereotype is received positively, as is sometimes the case with teasing (Keltner et al., 1998), nor does it account for any situations in which a positive stereotype is endorsed as in the provided example.

Figure 4.

Endorsing Meme (Panel A) and Subverting Meme (Panel B) Examples.



In contrast, subversive stereotype humor implicitly confronts or challenges the stereotype belief about the target group that is activated (Holmes & Marra, 2002). Subversive stereotype jokes not only activate the stereotype, but also provoke reevaluation of that stereotype, and/or the societal context in which it exists (Kramer, 2020). Consider the example in Panel B of Figure 3 (taken from Miller et al., 2019). The meme immediately establishes a conceptual framing surrounding the group (Black men) and the stereotype (Athletic) by posing the question "Why was the Black man good at basketball?" While it is never explicitly mentioned, the stereotype that Black men are naturally athletic is activated by content in the set up ("…Black man good at basketball"), particularly for those that have knowledge of this stereotype (Sailes, 1993). Activation of the stereotype offers one solution to this question in the viewer's mind. However, in the punchline "because he practiced", the viewer must consider this alternative explanation (which does not rely on and is completely independent of the stereotype) for basketball prowess in this individual, thus challenging the stereotype by invalidating its legitimacy in this humorous context.

Recent work has begun examining consequences of both jokes which endorse and subvert group stereotypes (Baumgartner & Morris, 2008; Saucier et al., 2018; Miller et al., 2019; Riquelme et al., 2021). For example, Riquelme and colleagues (2021) found that exposure to subversive anti-sexist humor, compared to neutral humor with no mention of gender, had positive implications for collective action for gender equality, particularly for individuals with weaker feminist identification. Similarly, Saucier and colleagues (2018) found that exposure to subversive racial humor (i.e., "What do you call a Black guy who flies a plane? A pilot you fucking racist!") led to lower expressions of

prejudice, operationalized as quantity of participant-reported negative versus positive stereotypes towards the target group, as compared to exposure to a disparaging joke (i.e., "Where do you hide your money from a Black thief? In your books."). While this work has focused on the beneficial downstream social outcomes of encountering subversive jokes, subversion is bistable and does necessarily come with the possibility of being misinterpreted as endorsing a stereotype (the "misunderstood"; Rappaport, 2005).

This misinterpretation has been documented in research examining reactions to the political comedy show The Colbert Report (Baumgartner & Morris, 2008). Even though Colbert satirized extreme right-wing views in an attempt to subvert conservative ideological arguments, participants who viewed The Colbert Report reported increased affinity towards conservatives and their policies. More recent studies have demonstrated comparable backfiring effects (Saucier et al., 2018); in this work, almost a third of participants interpreted researcher-selected subversive humor as endorsing negative stereotypes and discriminating against the targeted group. The same joke (i.e., "What do you call a Black guy who flies a plane?...") was used by Miller and colleagues (2019) in an experiment that investigated individual differences in perceptions of subversive versus disparaging racial humor. This package of studies examined tendencies to perceive racial humor (PMAPS), internal motivations to suppress prejudice (IMS), and modern racism (MRS); results showed that tendencies to both perceive and suppress racial prejudice were associated with more positive reactions to subversive humor. However, a subversion backfiring pattern emerged; participants higher in motivation to suppress prejudice found disparaging humor more racist, but these participants also found *subversive* humor more racist. Subsequent experiments using many instances of

experimenter-chosen subversive memes showed the same effect. Their work points to the complicated nature of both selecting and processing subversive humor.

Clearly, reactions to stereotype jokes are highly subjective and often independent of joker intent. As we see in both the Chris Rock example and published research, even when stereotype humor is shared with the intent to subvert stereotypes, there are individuals who interpret the same humor as endorsing stereotypes. Can we predict, based on some objective features of a joke, when an individual will perceive a stereotype joke as supportive/endorsing versus challenging/subversive? While the factors which influence this process have yet to be put under rigorous test, variations in how a stereotype is presented in a joke serve as likely candidates that may contribute to the differences in interpreting the subtext of stereotype humor.

Explicitness and Distortion

The current study examines two possibilities for objective features of a stereotype joke that would predict interpretations of endorsement versus subversion of the stereotype: 1) *level of explicitness of stereotyping* and 2) *degree of distortion of the stereotype*.

Explicitness of Stereotyping.

In some cases, jokes will explicitly mention the group and directly associate the group to their group stereotype. Two examples are offered in the first column of Table 1. In the first example (top left), the joker overtly mentions Mormons ("I was raised Mormon"; "Oh Billy you're Mormon") and goes on to directly tie this group to the stereotype that Mormons are polygamous ("How many wives do you have?"; "mom…s"). Similarly, in the next example (bottom left), the group being stereotyped in

the joke is overtly mentioned ("immigrants") and explicitly tied to their group stereotype ("do jobs white people don't want to do"). Thus, stereotype jokes that are explicit mention both the group and the group stereotype.

Table 1.

Jokes Highlighting Dimensions of Explicitness and Distortion.

	Explicit Association	**Implicit Association**
No Stereotype Distortion	"So, I was raised Mormon and whenever people found out I was Mormon as a kid they'd always tease me they'd be like "oh Billy you're Mormon, how many wives do you have?" and that hurt my feelings, I'd start to cry, and run to my mom…s"	"Police in Queens were searching for a suspect who hijacked a bus and then immediately crashed it into a utility pole. Okay, so we know it's a woman."
Stereotype Distortion	"They always say that immigrants do jobs that white people don't want to do. I must be white because I don't want to do that shit either."	"There were a lot of exchange students from Asia at my school. One gave me a ride to school one day. It was the most terrifying experience of my life. It got me thinking…Pearl Harbor might have been an accident"

In other cases, the group and the group stereotype are not both explicitly mentioned but become activated in our conscious awareness implicitly through spreading activation. Simply mentioning the group will elicit associated group characteristics (i.e., stereotypes). Several studies have demonstrated the implicit nature of stereotyping, showing it occurs when individuals aren't even aware of stereotype relevant information (Devine, 1989; Moskowitz, 2005) and when individuals try to ignore stereotypes or actively disavow them (Bargh et al., 1996; Galinsky & Moskowitz, 2000; Monin &

Miller, 2001). Based on theories of spreading activation, we can expect the reverse to be true; mentioning a group stereotype will bring to mind the associated group. Therefore, implicit stereotype jokes are those which mention only the group or only the associated stereotype, but not both.

Two examples of implicit stereotype humor are included in Table 1 (right column). In the first example (top right), only the group is mentioned (i.e., women). In order to understand the punchline ("Okay, so we know it's a woman"), the listener must call to mind the "bad driving stereotype" associated with women (Moè et al., 2015) themselves. This is also the case with the second example (bottom right). The group is explicitly mentioned ("students from Asia") to establish the subject of the joke. However, the stereotype ("Asians are bad drivers") is only ever alluded to ("terrifying experience" in reference to a ride to school; "Pearl Harbor might have been an accident"). These jokes are rather confusing if the stereotype is unknown.

Whether a stereotype is explicitly or implicitly tied to a group within a joke, the stereotype still becomes activated concurrently with the group by the end of the humor episode. The default perception from the mental coactivation of a group with their stereotype then should be an interpretation of endorsement (support or reinforcement of the association of the group with their stereotype). For this reason, it is also critical to consider whether, over the course of a joke, the stereotype is distorted in some way.

Stereotype Distortion.

There are instances in which a stereotype contained in a joke is activated in a listener's mind as the prototypical schematic representation ("at face value") and there are other cases in which a stereotype is manipulated in some way such that the schematic

representation is stretched or altered in the listeners mind. This "distortion" phenomenon is akin to the difference between looking at a mirror and seeing a true reflection versus looking in a funhouse or carnival mirror which deliberately warps the reflection of a viewer through compression or elongation.

Examples of stereotype jokes in which stereotypes are not distorted are presented in the top row of Table 1. The Mormon joke described earlier relies on the knowledge of the stereotype that Mormons have multiple wives. When one encounters the punchline ("run home to my mom…s"), the stereotype that Mormons are polygamous is not altered, it is simply activated at face value. This can also be seen in the bus crash joke. The punchline "Ok, so we know it's a woman" requires listeners to call to mind the "women are bad drivers" stereotype as-is in order to understand the joke.

In contrast, the stereotype jokes in the bottom row of Table 1 are those in which a stereotype is distorted in some way. The Immigrant joke explicitly associates the group ("immigrants") with the group stereotype ("do jobs white people don't want to"), but during the punchline ("I must be white . . .") the stereotype is countered. The comedian, who is himself a Mexican immigrant, negates this stereotype in the punchline by presenting himself as a counterexample (an immigrant is unwilling to fulfill the stereotype) thus invalidating the generalizability of the stereotyped information. The listener is left to either accept that the comedian has changed race (unlikely interpretation) or question the legitimacy of the stereotype in the joke. The Pearl Harbor joke offers an example of this in an implicit stereotype joke. The punchline "Pearl Harbor might have been an accident" distorts the stereotype by being exaggerated to a hyperbolic

degree: Asians are *such* bad drivers that Kamikaze pilots accidentally crashed into the American fleet of ships at Pearl Harbor.

Distortion can occur through a variety of ways as described in the previous examples. On one end, stereotypes can be extremely exaggerated or extrapolated to some impossible context; each of these processes stretches the schema of the stereotype outside of the typical realm in which the stereotype is encountered. On the other hand, stereotypes can experience a sort of compression through minimizing the stereotype or negating the stereotype by creating a situation in which the stereotype is put forth along with counterevidence. Distortion of a stereotype should be associated with greater probability of interpreting a stereotype joke as subversive because distortion inherently creates a discrepancy between the typical schematic representation of the stereotype and the distorted representation presented by the joke. This process is akin to mental accommodation (Piaget, 1970; Block, 1982) in which individuals experience the modification of a schema due to encountering and having to integrate information in the environment that conflicts with the schema.

How do these dimensions (explicitness, stereotype distortion) interact to predict interpretations of endorsement versus subversion of a stereotype in a stereotype joke? Predictions are visualized in Figure 5 below. As previously mentioned, the default assumption for most stereotype jokes is that they are endorsing the stereotype referenced in the joke due to the coactivation of the group and the group stereotype (Rappaport, 2005). Stereotype jokes in which the stereotype is not distorted (presented at-face-value) do not require listeners to challenge or question the viability of the stereotype and are

predicted to be interpreted primarily as endorsement, regardless of whether the stereotype is explicitly or implicitly linked to the group.

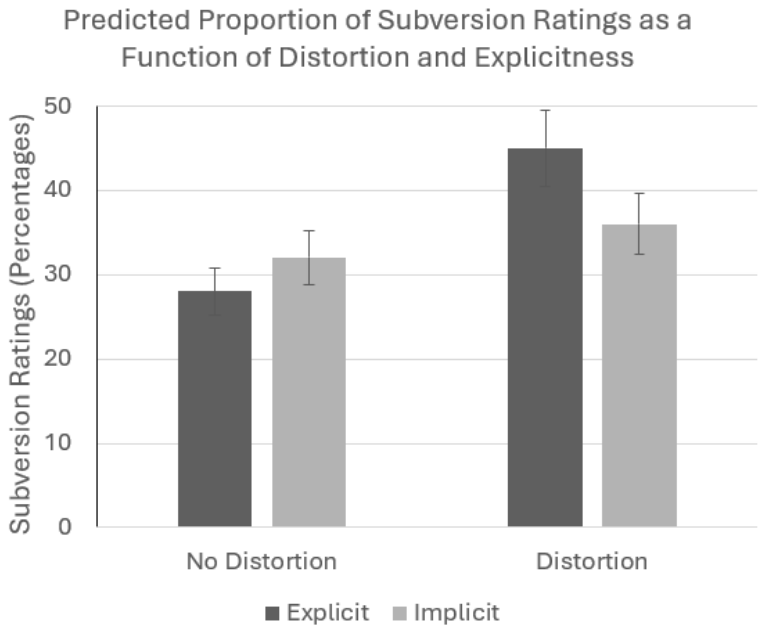
While distorting a stereotype in some way is more in line with subversion than endorsement, I predict an interesting interaction when considering the effects of this factor in the context of the explicitness dimension. In the case of explicit jokes that have been distorted, the listener is directly presented with stereotyped information about a specific group by another individual and faces a discrepancy between the stereotype as they typically schematically represent it and the distorted version created through humor presented to the listener by the joker. The schema discrepancy created through explicit distortion should then heighten the likelihood that a listener infers stereotype subversion, (see Figure 4). However, these types of jokes may still be interpreted as endorsement if the distorted version of the stereotype is not comprehended, the listener believes the stereotype more rigidly and is unwilling to question it, or the listener is more sensitive to prejudice (Miller et al., 2019).

The final endorsement versus subversion prediction concerns implicit stereotype jokes in which the stereotype is distorted. I predict that, although the stereotype is distorted in some way, individuals will still primarily interpret implicit jokes as endorsement. When a joke implicitly associates a group with their stereotype, the effect of distorting the stereotype is diminished because the listener has to rely on spreading activation to call to mind the stereotype as-is in order to comprehend the joke. Individuals are biased to favor ideas or information that is perceived as belonging to them (ownership bias; Van Swol & Sniezek, 2002; Yaniv & Kleinberger, 2000). Since implicit jokes require listeners to activate the stereotype schema themselves (rather than having it

presented to them in the case of explicit jokes), this personally activated schema should hold more weight when interpreting subtext, making it more difficult to interpret the distorted schema created by the joker as challenging the stereotype. This should lead to the interpretation of the stereotype as being upheld or endorsed within the joke.

Figure 5.

Predicted Proportion of Subversion Ratings.



Note. No instance of stereotype humor is expected to be interpreted by 100% of viewers as endorsement, therefore base subversion proportion for non-distorted stereotype jokes is expected to hang relatively low at 30% (70% interpretation of endorsement).

The Current Study

The current study investigates the interactive effect of i) explicitness of stereotyping and ii) stereotype distortion on interpretations of subtext (endorsement versus subversion) in several stereotype memes. Four groups (Asian, Hispanic, Irish, White) and associated group stereotypes are used in this study. Additionally, I am

interested in how explicitness and distortion are related to judgments of a joke's funniness and offensiveness. Here I employ a 2 (explicit/implicit) x 2 (no stereotype distortion/stereotype distortion) completely within-subjects design in which participants first categorize memes as endorsing or subverting the stereotype. This categorization serves as the primary binary dependent variable. Participants then review the memes again and provide ratings of funniness and offensiveness for each meme. Joke funniness and offensiveness ratings will be treated as continuous variables and serve as additional dependent variables. Main effects and moderation by target group will also be explored.

Hypotheses for the primary outcome of subversion versus endorsement are as follows: an interaction between distortion and explicitness such that explicit memes that are distorted will result in significantly greater interpretations of subversion (H1a), a main effect of stereotype distortion such that distorted memes will result in significantly greater judgments of subversion (H1b), and no effect of explicitness on subversion ratings (H1c). In terms of funniness ratings, hypotheses are as follows: an interaction between distortion and explicitness such that explicit memes that are distorted will be rated significantly funnier (H2a), a main effect of stereotype distortion such that distorted memes will be rated significantly funnier (H2b), and a no main effect of explicitness on funny ratings (H2c). Finally, in terms of offense ratings, hypotheses are as follows: an interaction between distortion and explicitness such that explicit distortion memes will be rated significantly less offensive (H3a), a main effect of distortion such that distorted memes will be rated significantly less offensive (H3b), and a main effect of explicitness such that explicit jokes will be rated significantly more offensive (H3c).

CHAPTER 3

METHODS

This study was pre-registered. I report all manipulations, measures, and exclusions in these studies. Materials, anonymized data, and analysis code for all studies can be found at https://osf.io/jas6x/.

Participants

Based on an a priori power analysis, we aimed for a minimum final sample of N = 116 participants for this wholly within-subjects experiment; this sample size provides 99% power to detect an effect size of partial eta squared = .01, assuming alpha = .05. (calculated in GPower—ANOVA). Because this is the first study to examine these specific factors in the context of stereotype humor, I chose to use the smallest eta squared of interest (.01) and an exceedingly high threshold for power (.99) to ensure the sample is sufficiently powered to detect even tiny effects. However, primary analyses will rely on multilevel binary logistic regression to examine the probability of endorsement versus subversion and the above power analysis does not account for clustering of responses at the participant level. Work by Peduzzi et al. (1996) outlines guidelines for minimum sample size calculation for multilevel logistic regression and suggests that a minimum sample size of 100 participants is necessary.

Participants were recruited online using the ASU Psychology Subject Research Pool and received one hour of course credit for completing the study. In total, 131 participants completed the study, 27 participants were removed for failing attention checks, failing knowledge checks, or insufficient English fluency. This left us with a final sample of 104 participants. A post hoc power analysis in GPower revealed that our final

N = 104 sample had greater than 98% power to detect an effect size of eta squared = .01. The mean age of participants was 19.59 (2.62) years, with a range of 18 to 40 years old. Participants were 60.6% female; 65.4% White, 29.8% Hispanic, 3.8% Black/African American, 2.9% East Asian, 6.7% Southeast Asian, 5.8% South Asian, 1% Pacific Islander, 4.8% Middle eastern, 1.9% Native American, and 1.9% other ethnicity. Participants could select more than one race/ethnicity and 21.2% reported being of mixed race or ethnic background. Of the final sample, 97.1% of participants reported at least occasionally encountering humor about stereotypes in their everyday lives, with 34.6% reporting frequent encounters, and 32.7% reporting very frequent encounters with stereotype humor.

Procedure

Participants first read over a consent form and consented to participate in the study. Next, participants indicated familiarity with 15 different group stereotypes, four of which were the target stereotypes for analyses. After these ratings, participants were provided definitions for the two key concepts required for the main task in the study, endorsement (support/uphold) and subversion (challenge/undermine); each definition was presented on the screen for 20 seconds before participants could move on to ensure definitions were read thoroughly. Two randomly presented multiple-choice questions assessed participants' comprehension of endorsement and subversion and served as knowledge checks.

Participants were randomly shown the 16 stereotype memes, one at a time, and asked to categorize each as either endorsing (upholding/supporting) or subverting (undermining/challenging) the stereotype contained within the meme. Participants could

take as much time as needed to make their decision; time to make each decision was recorded but was not used in current analyses. An additional meme with text directing participants to select "undermines/challenges" was randomly embedded in this task and served as an attention check for this portion of the experiment.

Following the endorsement vs. subversion subtext categorization task, participants were told that they would rate the memes they had previously seen for how funny and offensive they found each meme. Before the rating task, participants were instructed that ratings should be made based on how personally funny and offensive they found each meme, rather than basing ratings on how someone else or society at large would rate the memes. A multiple-choice question asking participants to report that ratings would be made based on personal judgments was used during this introduction to the task as an attention check.

Next, participants were randomly presented each meme one at a time and instructed to rate them for how funny and how offensive each meme was to them; funny and offensive ratings were asked separately, the meme was displayed above each question, and presentation of these two questions was randomized. Once these two questions were completed, the meme was presented again, and participants were asked to report which stereotype the meme was referencing using a multiple-choice question including the correct stereotype and three alternative incorrect options. This question was used as an additional attention check. Once all funny and offense ratings were completed, participants were asked to report how generally offensive they found each of the four target stereotypes. Each stereotype was presented one at a time and in random order.

After all tasks relevant to the stereotypes and memes were done, participants completed measures of divergent thinking, humor styles, and motivations to control prejudice. These three measures were presented to participants in random order. Only scores on motivations to control prejudice are included in exploratory analyses; divergent thinking and humor style will be examined in the future. Following this, participants completed demographics and reported whether they identified with any of the target groups presented in the memes, as well as how often they encounter humor about stereotypes in their day to day lives. Finally, participants reviewed a debriefing statement and clicked a link directing them to an external survey which prompted them to fill out their information to receive credit for their participation.

Materials and Measures

Stereotype Familiarity Ratings.

Participants rated familiarity with 15 stereotypes about different groups (including the 4 target groups) on a Likert scale from 1 ("Not at all familiar/Never heard of it") to 5 ("Extremely familiar"). Familiarity ratings were conducted at the beginning of the experiment as a covert way to ensure that all participants would have some exposure to the four target stereotypes prior to the main task.

Memes.

Four groups (and associated group stereotypes) were selected for this study, including Asian ("Asian parents want their kids to be doctors, above all else"), Hispanic ("Hispanic mothers threaten corporal punishment with a sandal ("chancla") for misbehavior"), Irish ("Irish people drink a lot"), and White ("White people are bad at dancing"). Stereotypes were selected for being relatively well-known or statistically

prevalent and slightly negative but not extremely derogative. Only one group was mentioned in each meme and the same stereotype was referenced across each group's set of memes. Memes were created using the website https://imgflip.com/memegenerator.

Each group was referenced in four memes, which systematically varied by the two factors of interest (explicitness, distortion), and therefore 16 memes served as target stimuli. Explicit memes contained text pertaining to both the group and their stereotype; implicit memes contained text only pertaining to the group (implying the stereotype) or text only pertaining to the stereotype (implying the group). Non-distorted memes presented the stereotype "as-is" or at face value; distorted memes present the stereotype through exaggeration or over extrapolation of the stereotype to some other ridiculous context. While distortion can occur through exaggeration, extrapolation, minimizing, and counter evidence, only exaggeration and extrapolation were used in the current study.

Endorsement vs. Subversion Categorization Task.

Participants were asked to categorize each of the 16 stereotype memes as either endorsing or subverting the stereotype contained within the meme. Participants were required to fill in the following sentence with one of two options: "This meme _____ the stereotype contained in it." Endorsement ratings were coded as 0, and subversion ratings were coded as 1.

In the task, endorsement was labeled "Upholds/Supports", and the following definition was provided to participants: "In this context, the verb uphold means to support or endorse belief in a stereotype by assuming its validity, truth, or integrity. When you uphold something, you are ensuring that it remains in place or is preserved without weakening or compromise. It implies keeping something in a particular state or condition.

Memes that uphold/support the stereotype they are about take for granted that the stereotype is actually true or does apply to the group it is about. Basically, the meme reinforces the stereotype."

Subversion was labeled "Undermines/Challenges", and the following definition was provided to participants: "In this context, the verb undermine means to challenge or subvert belief in a stereotype by gradually eroding the foundation of its validity, truth, or integrity. When you undermine something, you are challenging its credibility with the intent to cause doubt and weaken support for its reliability. It implies a desire to cause significant change or disruption in a particular state or condition. Memes that undermine/challenge the stereotype they are about make the viewer feel like the stereotype is not accurate or does not really apply to all members of the group it is about. Basically, the meme is making fun of the stereotype or making the stereotype seem ridiculous.

Funny, Offensive, and General Stereotype Offense Rating Tasks.

Funny and Offensive ratings for each individual meme were assessed through separate 6-point Likert scales ranging from 0 ("Not at all funny"/ "Not at all offensive") to 5 ("Funniest meme I have ever seen"/ "Most offensive meme I have ever seen"). The general offensiveness of the four stereotypes was assessed through a 5-point Likert scale ranging from 1 ("Not at all offensive") to 5 ("Extremely offensive").

Motivations to Control Prejudicial Responses – Concern Subscale.

This scale taps into differences in motivation to control expressions of prejudice. Based on a confirmatory factor analysis of the original Motivation to Control Prejudiced Reactions Scale (Morrison et al., 2014), seven items surrounding personal concern over

appearing prejudiced were selected from the 17 items that make up the original Dunton and Fazio (1997) MCPR measure. Example items include: “In today’s society it is important that one not be perceived as prejudiced in any manner” and “I get angry with myself when I have a thought or feeling that might be considered prejudiced.”

Humor Styles Questionnaire.

The Humor Styles Questionnaire is a 32-item survey (8 items per subscale) designed to assess variation in individual differences in sense of humor (how often people laugh and appreciate jokes/humor) and specific dimensions or styles of humor (Martin et al., 2003). This measure was assessed but not used for current analyses.

Divergent Association Task.

This task serves as a new brief, reliable, and objective measure of divergent thinking (Olson et al., 2021). Participants named 10 words as different from each other as possible. Average semantic distance between the words is then calculated as a measure of divergent thinking capacity, but this measure is not used for current analyses.

Demographics.

Participants reported gender (man, woman, non-binary, decline to answer), age (free response text input), race/ethnicity (check all that apply), and English proficiency (ranging from “I have very limited English skills or none at all” to “I am a native English speaker”). Participants also reported identification with any of the target groups (check all that apply), as well as the frequency with which they encounter stereotype humor in their everyday lives on a 5-point Likert scale ranging from 0 (Never) to 4 (Very frequently).

Statistical Analyses

Primary analyses involve a series of multilevel repeated measures logistic regression models to determine the probability of a subversion rating (1), compared to endorsement rating (0), as a function of the two independent variables (explicitness, distortion) and their interaction (explicitness x distortion). These models accounted for the nesting of endorsement/subversion ratings (Level 1) in participants (Level 2) using a random intercept term. Group was also examined as a covariate and moderator to see whether ratings of endorsement/subversion vary by group stereotype. Two additional series of multilevel regression models were used to investigate the effect of explicitness and distortion and their interaction on funniness and offensiveness ratings. These models also accounted for the nesting of funny and offense ratings (Level 1) in participants (Level 2) using a random intercept term. Group was added to these models to see if ratings of these constructs vary by group stereotype as well. Concern with appearing prejudiced was included in subsequent exploratory models for each dependent variable.

To test the effect of our factors (explicitness, distortion) and the group variable (Asian, Hispanic, Irish, White) on the outcomes of interest (subversion, funny, offense), I conducted a series of chi-square fit tests between 1) the base model (only the random factor), 2) the model that included the random factor and just explicitness, distortion, and the interaction, and 3) the model that included these terms interacting with the group variable. If the model fit test was statistically significant, it indicated that the more complex model—the one including group—was a better fit for the data. Only when the group variable significantly improved fit was it considered a moderator.

CHAPTER 4

RESULTS

Analyses were conducted in R version [4.2.3] (R Core Team, 2023); mixed effects models were fitted using the lme4 package (Bates, 2010). Funny ratings underwent square root transformation to normalize the distribution and centering before running the models. Offense ratings were heavily positively skewed due to a very large number of zeros, and therefore, multilevel negative binomial regression was used to account for the over-dispersion in this outcome variable. All models account for individual differences by including participant ID as a random factor.

Descriptive Statistics

Subversion rating frequencies, means and standard deviations for funny and means and standard deviations for offense ratings are provided for each individual meme as well as by meme group (set of four averaged), presented in Table 2. Also included are the percent of participants that identified with each group, means and standard deviations for participant familiarity with each group stereotype, and means and standard deviations for general offense participants felt towards each group stereotype.

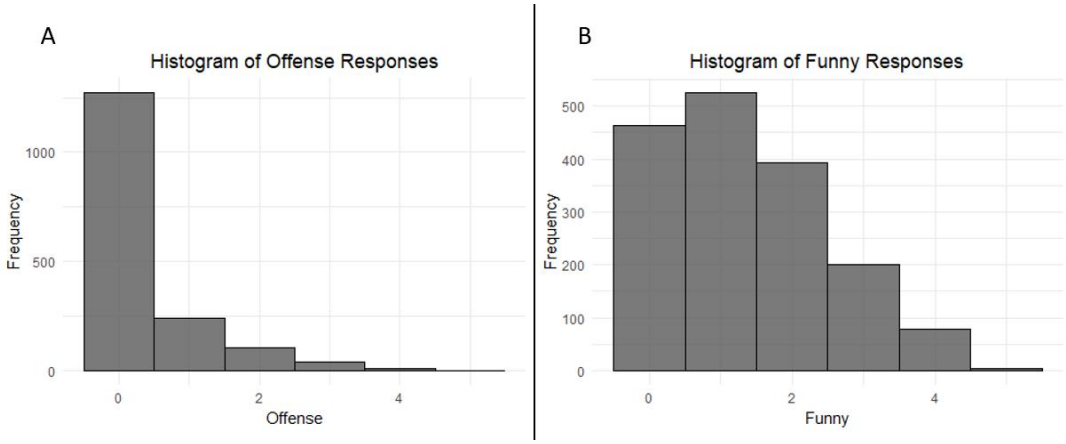
Approximately two-thirds of the sample identified as White, making it the largest represented group of the four target groups, followed by Hispanic (almost one-third), Asian (almost one-fifth), and the least number of participants identified as Irish (one-eighth). All stereotypes were relatively well-known to the sample of participants, with familiarity averages of each stereotype greater than the midpoint of the scale ("somewhat familiar"). All stereotypes hung around slightly to moderately offensive; the Asian stereotype ("demanding parents wanting their kids to be doctors") was reported most

offensive of the four and the White stereotype ("can't dance") was reported least offensive; however, these differences were not significant.

All memes were, on average, extremely low in offense and slightly to moderately funny (see Figure 6 below); over 75% of meme offense responses fell into the "not at all offensive" category and over 72% of meme funny ratings were categorized as at least slightly funny. Participants rated memes about the White stereotype as least offensive and funniest, while memes about the Asian stereotype were most offensive and least funny. This pattern in meme offense ratings matches participant reports of general stereotype offense level. Unlike previous work relying on highly disparaging stereotypes (with frequent use of racial slurs; Miller, 2019), the memes in this study allowed for testing key hypotheses in cases where stereotype humor was viewed as less outright derogatory. This type of stereotype humor is presumably more common than highly disparaging humor.

Figure 6.

Histogram of Meme Offense Ratings (Panel A) and Meme Funniness Ratings (Panel B).



Note. Frequency ranges are on different scales for each type of response. Both histograms show all participant responses for funny and offense ratings of all memes.

Table 2.

Descriptive Statistics for Stereotype Variables and Memes.

Group	Stereotype Statistics	Percent Subversion	Funny Ratings	Offense Ratings
Irish				
% Identify	12.5%	35.7%	1.44 (0.88)	0.36 (0.66)
Familiarity	3.5 (1.3)			
Offense	1.48 (0.8)			
	Meme 1 (Exp_ND)	31.7%	1.6 (1.15)	0.28 (0.69)
	Meme 2 (Imp_ND)	26.9%	1.29 (1.06)	0.24 (0.6)
	Meme 3 (Exp_D)	42.3%	1.47 (1.18)	0.42 (0.83)
	Meme 4 (Imp_D)	41.3%	1.39 (1.2)	0.48 (0.85)
White				
% Identify	65.4%	32.9%	1.46 (0.96)	0.26 (0.53)
Familiarity	3.73 (1.28)			
Offense	1.37 (0.65)			
	Meme 1 (Exp_ND)	29.8%	1.28 (1.16)	0.19 (0.5)
	Meme 2 (Imp_ND)	31.7%	1.79 (1.22)	0.29 (0.62)
	Meme 3 (Exp_D)	36.5%	1.28 (1.32)	0.28 (0.62)
	Meme 4 (Imp_D)	33.7%	1.49 (1.2)	0.28 (0.72)
Asian				
% Identify	18.3%	28.4%	1.25 (0.90)	0.48 (0.71)
Familiarity	4.23 (1.01)			
Offense	1.67 (0.9)			
	Meme 1 (Exp_ND)	27.9%	1.53 (1.1)	0.43 (0.71)
	Meme 2 (Imp_ND)	13.5%	1.04 (1.03)	0.46 (0.76)
	Meme 3 (Exp_D)	39.4%	1.18 (1.12)	0.52 (0.87)
	Meme 4 (Imp_D)	32.7%	1.26 (1.17)	0.5 (0.82)
Hispanic				
% Identify	29.8%	26.4%	1.26 (0.84)	0.35 (0.70)
Familiarity	3.82 (1.41)			
Offense	1.51 (0.87)			
	Meme 1 (Exp_ND)	20.2%	1.01 (1.05)	0.38 (0.83)
	Meme 2 (Imp_ND)	9.6%	1.4 (1.17)	0.34 (0.9)
	Meme 3 (Exp_D)	38.5%	1.29 (1.12)	0.46 (0.89)
	Meme 4 (Imp_D)	37.5%	1.33 (1.1)	0.24 (0.63)

Note. Descriptive statistics for participant stereotype identification (percentage); stereotype familiarity (*M* and *SD*); stereotype offense (*M* and *SD*). The following statistics are provided for each meme individually and the total statistics by group: percent of participants that selected subversion; funny ratings (*M* and *SD*); offense ratings (*M* and *SD*). Exp = Explicit; Imp = Implicit; ND = No Distortion; D = Distortion.

Table 3.

Means, Standard Deviations, and Correlations for Study Variables.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Total Subversion Ratings (%)	31.37	24.69	—									
2. MCPR Concern	5.70	8.16	-0.05	—								
3. Offensive (Irish)	0.36	0.66	0.05	**0.36****	—							
4. Offensive (White)	0.26	0.53	0.18	**0.37****	**0.72****	—						
5. Offensive (Asian)	0.48	0.71	0.03	**0.27****	**0.53****	**0.42****	—					
6. Offensive (Hispanic)	0.35	0.70	0.05	**0.24****	**0.65****	**0.46****	**0.73****	—				
7. Funny (Irish)	1.44	0.88	0.06	-0.08	**-0.23***	-.010	-0.19	-0.14	—			
8. Funny (White)	1.46	0.96	0.01	**0.22***	0.07	0.02	-0.09	0.02	**0.60****	—		
9. Funny (Asian)	1.25	0.90	-0.03	-0.01	-0.11	-0.08	**-0.30****	-0.08	**0.55****	**0.60****	—	
10. Funny (Hispanic)	1.26	0.84	0.13	0.02	0.00	0.03	**-0.25***	**-0.28****	**0.49****	**0.59****	**0.61****	—

Note. $N = 104$. *M* and *SD* are used to represent mean and standard deviation.* indicates $p < .05$. ** indicates $p < .01$. Correlations of offense and funny ratings across groups are highlighted in grey, those between offense and funny ratings in light grey.

Correlations

Correlations between participant total subversion percentage, concern over appearing prejudiced, meme offense ratings (by group), and meme funny ratings (by group) are presented in Table 3. The total proportion of subversion ratings made by each participant did not correlate with any other variable. Participant concern over appearing prejudiced significantly positively correlated with offense ratings for each group's set of memes, as well as funny ratings only for memes about the White stereotype; all correlations were weak. Higher levels of concern over appearing prejudiced was associated with higher offense ratings on all memes.

Two key clusters of correlations emerged, one around offense ratings and the other around funny ratings for memes across the different groups. Offense ratings positively correlated significantly among all groups, ranging from moderate to strong associations; the strongest offense correlations were between Hispanic and Asian memes ($r = 0.73$) followed by Irish and White memes ($r = 0.72$). Similarly, funny ratings positively correlated significantly among all groups, all to a moderate degree; the weakest funny correlation was between Irish and Hispanic memes ($r = 0.49$). Overall, across the groups, participant ratings of offensiveness and funniness were positively associated. Additionally, each group's funny ratings negatively, but rather weakly, correlated with the same group's offense ratings. The exception to this pattern was the White memes, which did not demonstrate any correlation between funny and offense ratings. Finally, funny ratings for Hispanic memes significantly negatively correlated with offense ratings for Asian memes, and this correlation was also relatively weak ($r = -0.25$). This finding indicates perceptions of funny and offense are negatively related only for

underrepresented minority groups, and that the effect is stronger yet more generalized across the highly marginalized groups.

Results of Model Predicting Subversion

The model predicting subversion from distortion, explicitness, the distortion x explicitness interaction, group category, and participant ID (random intercept) performed significantly better than a) the base model which only included the random intercept (χ^2 = 72.59, $p < .001$), and b) the model that included all fixed factors but without group included ($\chi^2 = 14.15$, $p < .01$). Results of this model are presented in Table 4. Two responses from separate participants (one for Irish, one for Asian) were omitted because participants reported the wrong stereotype as present in the meme. Figure 7 shows distributions of actual endorsement and subversion ratings for all memes by target group.

Figure 7.

Bar Graphs of Meme Subtext Judgments by Group.

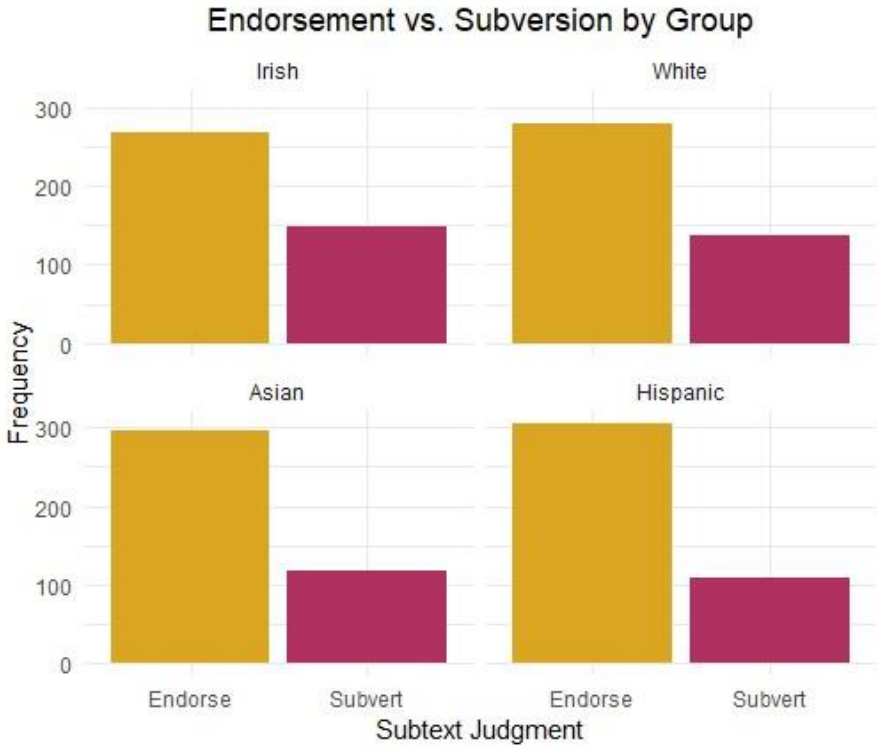


Table 4.

Multilevel Binary Logistic Regression Predicting Subversion Ratings

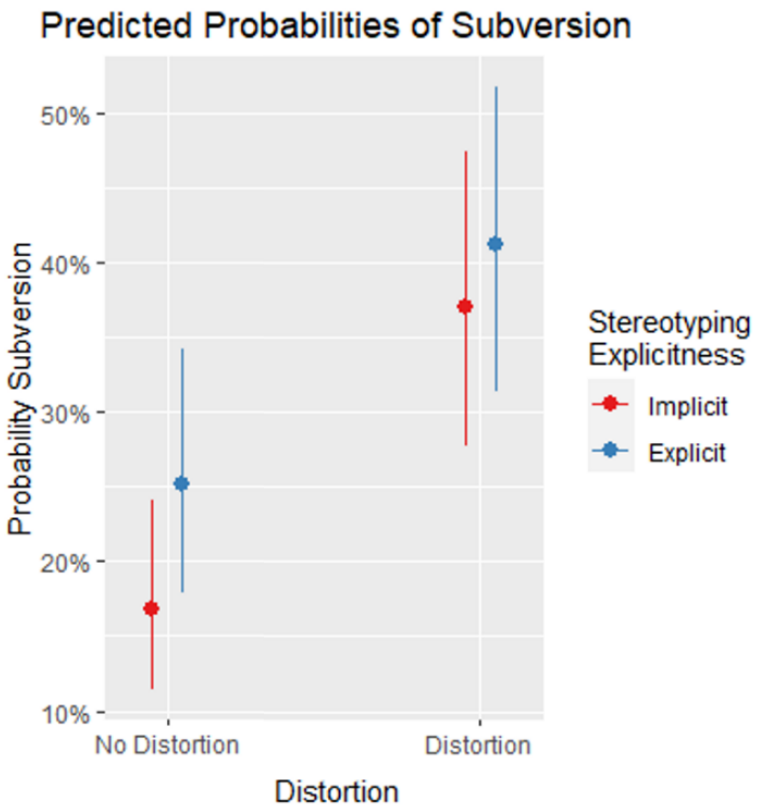
	Subversion			
Predictors	*Odds Ratios*	*std. Beta*	*CI*	*p*
(Intercept)	0.20	0.20	0.13 – 0.32	**<0.001**
Distortion	2.91	2.91	2.03 – 4.15	**<0.001**
Explicitness	1.66	1.66	1.15 – 2.38	**0.006**
Group [White]	0.85	0.85	0.61 – 1.18	0.330
Group [Asian]	0.63	0.63	0.45 – 0.89	**0.008**
Group [Hispanic]	0.55	0.55	0.39 – 0.78	**0.001**
Distortion × Explicitness	0.72	0.72	0.44 – 1.17	0.181
Random Effects				
σ^2	3.29			
$\tau_{00\ ID}$	2.23			
ICC	0.40			
N_{ID}	104			
Observations	1662			
Marginal R^2 / Conditional R^2	0.051 / 0.434			
AIC	1745.95			

Note. Irish is the reference group.

Results of the multilevel binary logistic regression show significant effects of both distortion and explicitness (see Figure 8). There was no significant interaction between explicitness and distortion, so Hypothesis 1a was not supported. Distorted memes were significantly more likely to be rated as subversive, providing evidence for Hypothesis 1b. Contrary to Hypothesis 1c, explicit memes were also more likely to be rated as subversive compared to implicit memes.

Figure 8.

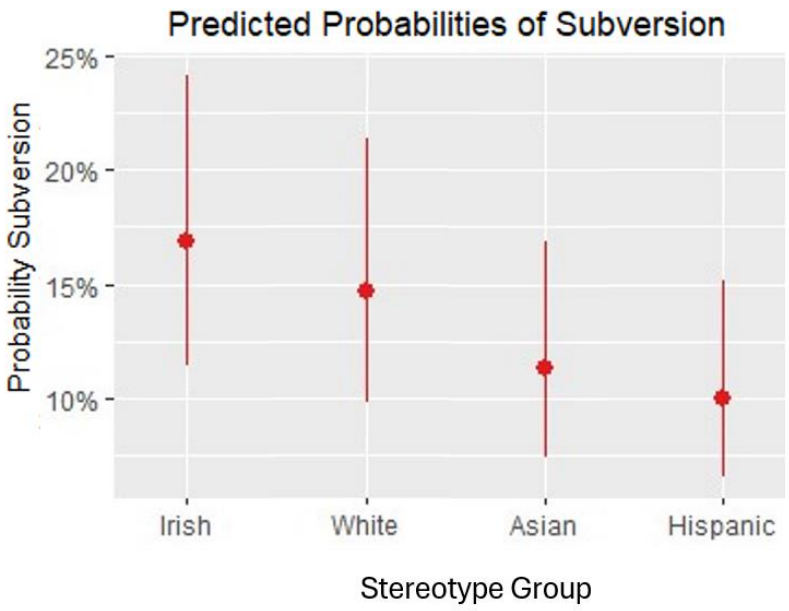
Interaction Plot of Predicted Subversion Probabilities.



Results also showed that 40% of the variability in subversion ratings was attributed to individual differences. An interesting main effect of meme group category emerged (see Figure 9). Irish memes were most likely to be rated subversive and served as the reference group. Memes about the Asian stereotype and the Hispanic stereotype ("use of chancla [sandal] for corporal punishment") were significantly less likely to be rated as subversive compared to memes about the Irish stereotype ("drink a lot"). These memes were also less likely to be rated subversive compared to the memes about the White stereotype, although this difference was not found to be significant.

Figure 9.

Effect of Stereotype Group on Subversion Probabilities



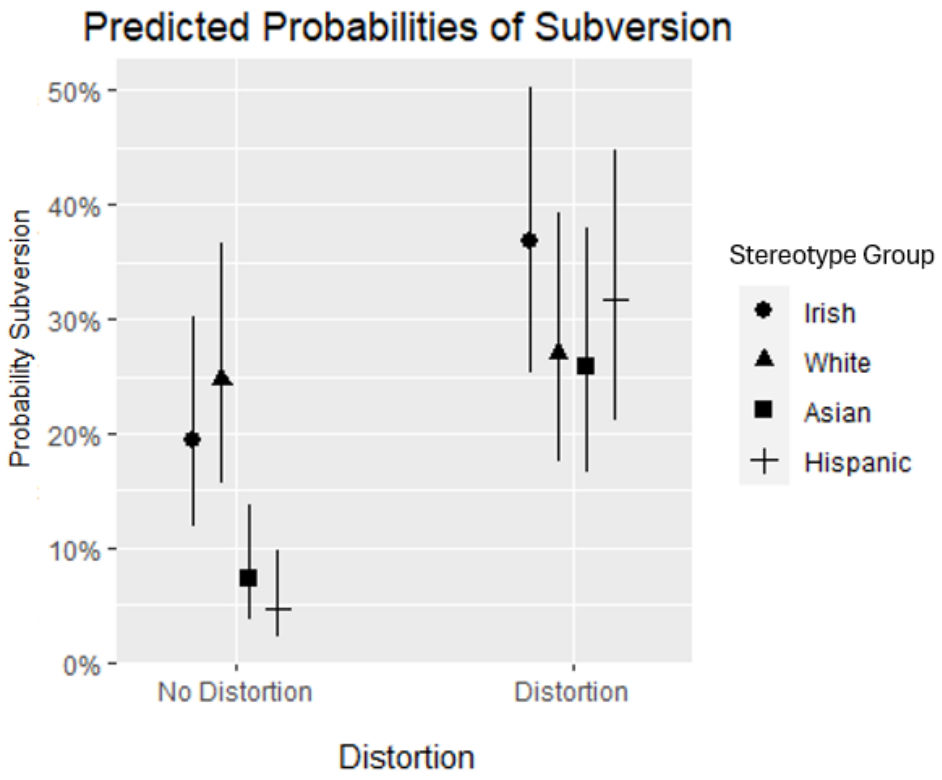
Exploratory Moderation Analyses (Subversion Ratings)

Additional exploratory analyses were conducted to investigate the moderating effect of group stereotype on ratings of meme subtext; while the sample was under-powered to detect moderating effects, I report the odds ratios here to compare effects sizes across groups. First, an additional model was run including interaction terms between target group and both explicitness and distortion, as well as their interaction. This model showed significantly better fit than the model including group as an additional predictor ($\chi^2 = 24.00$, $p = .013$). In this model, the effect of distortion remained stable, but did weaken slightly ($OR = 2.43$, $p = .01$). Asian ($OR = 0.33$, $p < .01$) and Hispanic ($OR = 0.21$, $p < .001$) memes were still significantly less likely to result in subversion ratings as well. The only significant moderation by group that emerged was on the effect of distortion in the case of Hispanic memes (see Figure 10). Distortion was

far more likely to result in subversion ratings for Hispanic memes (*OR* = 3.85, *p* = .015) as compared to the Irish (*OR* = 0.21, *p* > .05) and White memes (*OR* = 0.21, *p* > .05). While the moderation effect of group on distortion did not reach significance for the Asian memes (*OR* = 1.81, *p* > .05), a similar pattern to Hispanic memes emerged.

Figure 10.

Moderation of Distortion by Stereotype Group on Subversion Ratings



Next, analyses were run separately for each target group and an additional predictor indicating whether participants reported being in-group or outgroup with respect to the target group was included in the models predicting subversion ratings. These analyses were conducted to explore which groups were driving the effects in the main analyses. Table 5 summarizes effects for each analysis by target group. Distortion, explicitness, and their interaction were the primary predictors as in previous models. In-

Table 5.

Multilevel Binary Logistic Regression Models Predicting Subversion Ratings by Group.

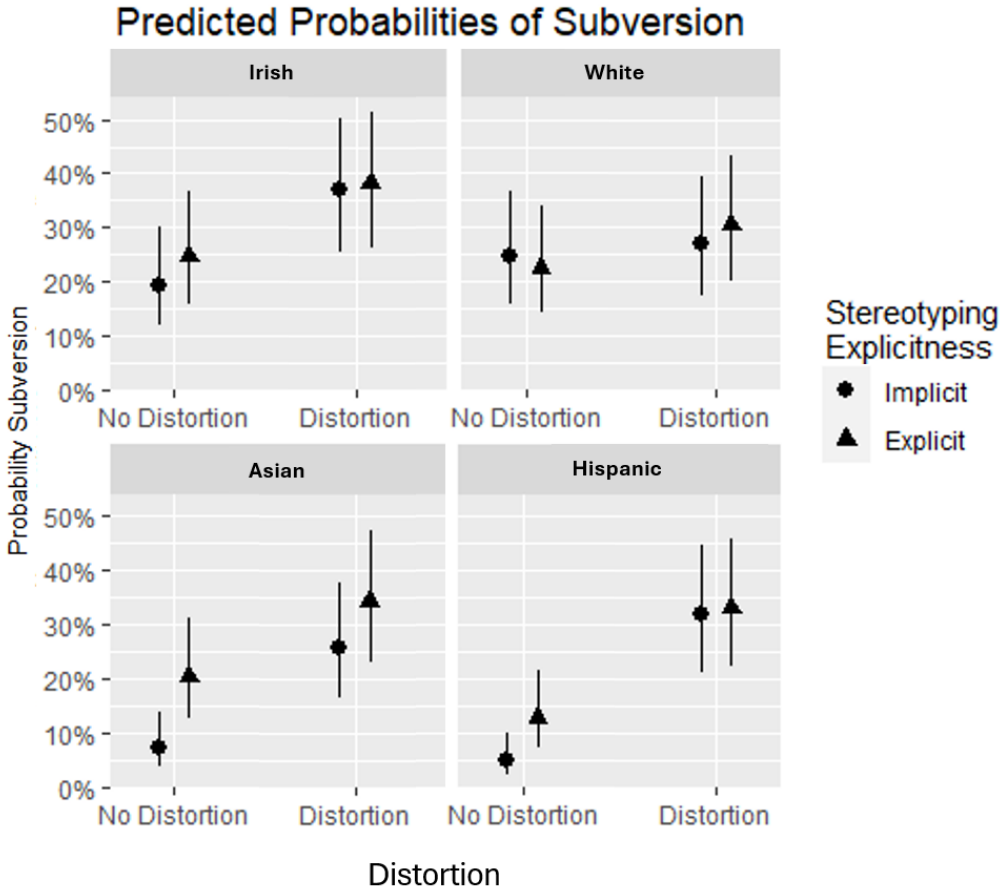
	Subversion			
	Irish	**White**	**Asian**	**Hispanic**
Predictors	*Odds Ratio*	*Odds Ratio*	*Odds Ratio*	*Odds Ratio*
Distortion	**2.97****	1.19	**4.68****	**8.86*****
Explicitness	1.50	0.57	**2.72***	**2.93***
In-Group	0.26	0.86	0.14	0.50
Distortion× Explicitness	0.62	1.47	0.67	0.37
Distortion × In-Group	0.75	0.93	5.60	1.64
Explicitness × In-Group	0.67	1.93	14.58^{+}	1.19
Distortion × Explicitness × In-Group	5.47	0.89	0.04^{+}	0.77
Random Effects				
σ^2	3.29	3.29	3.29	3.29
$\tau_{00\ ID}$	3.74	2.20	4.05	2.20
ICC	0.53	0.40	0.55	0.40
$N_{\ ID}$	104	104	104	104
Observations	415	416	415	416
Marginal R^2/ Conditional R^2	0.059 / 0.560	0.009 / 0.407	0.115 / 0.604	0.155 / 0.494
AIC	485.16	506.29	438.16	435.73

Note. $^{+}$ indicates $p < .10$.* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

group status of participants was also entered into the separate models, as well as interaction terms between in-group status and the previous terms. Effects are visualized in Figure 11 below.

Figure 11.

Interaction Plot of Predicted Subversion Probabilities by Group.



For the White memes, none of the variables were significant; subversion ratings were not predicted by any variables of interest for the White memes. For the Irish memes, only distortion showed a significant effect on subversion ratings ($OR = 2.97$, $p < .01$); distorted Irish memes were far more likely to result in subversion ratings compared to non-distorted Irish memes. For Asian memes, both distortion ($OR = 4.68$, $p = .001$) and

explicitness (OR = 2.72, p = .039) showed significant effects on subversion ratings; distorted memes and explicit memes were more likely to be rated as subversive compared to non-distorted memes and implicit memes, respectively. This result matches the effect pattern of the primary analysis predicting subversion conducted on all memes.

Similarly, for Hispanic memes, both distortion (OR = 8.86, p < .001) and explicitness (OR = 2.93, p = .047) showed significant effects on subversion ratings; distorted memes and explicit memes were more likely to be rated as subversive compared to non-distorted memes and implicit memes, respectively. In-group status did not show any significant effects across any of the groups. In the case of Asian memes, there were a few interactions with in-group status that were marginally significant, but analyses were underpowered to detect moderation effects, particularly for groups that had very few in-group members.

Results of Model Predicting Funny Ratings

Next, I present analyses predicting funny ratings (square root transformed). The model predicting funny ratings from distortion, explicitness, the distort x explicit interaction, and participant ID (random intercept) performed worse than the base model with only participant ID entered as a random factor; this difference was not significant based on the chi-square test. Adding group as a covariate to the model with all predictors showed a significantly better fit (χ^2 = 14.55, p < .01). In this full model, only the group variable showed a significant main effect, and no other predictor was significant. A model was fitted with only the group variable as a predictor and this model showed a better fit compared to the full model; since the chi-square test of these models was not significant, results are presented for the simpler model including only group.

Table 6.

Multilevel Regression Predicting Funny Ratings.

Funny Ratings (Square Root Transformed)				
Predictors	*Estimates*	*std. Beta*	*standardized CI*	*p*
(Intercept)	0.06	0.09	-0.05 – 0.23	0.200
Group [White]	-0.02	-0.03	-0.14 – 0.08	0.631
Group [Asian]	-0.11	-0.16	-0.27 – -0.05	**0.004**
Group [Hispanic]	-0.11	-0.17	-0.28 – -0.06	**0.003**
Random Effects				
σ^2	0.29			
$\tau_{00\ ID}$	0.15			
ICC	0.34			
N_{ID}	104			
Observations	1664			
Marginal R^2 / Conditional R^2	0.006 / 0.340			
AIC	2903.87			

Note. Irish is the reference group.

Results for the multilevel regression model predicting square root transformed funny ratings are presented in Table 6. Contrary to what was predicted, no support was found for Hypothesis 2a (interaction) or Hypothesis 2b (distortion). As there was no effect of explicitness on funny ratings, Hypothesis 2c was supported. Funny ratings were not dependent on stereotype distortion or explicitness of stereotyping (see Figure 12); target group ended up being the only significant predictor, with the two minoritized (Asian, Hispanic) groups' memes associated with significantly lower funny ratings compared to Irish and White memes.

Figure 12.

Interaction Plot of Predicted Funny Ratings.

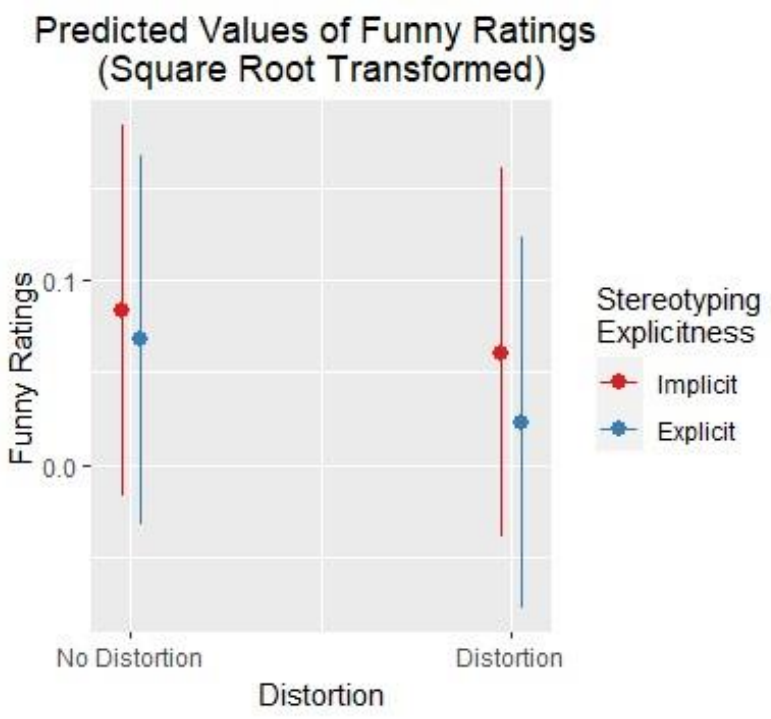
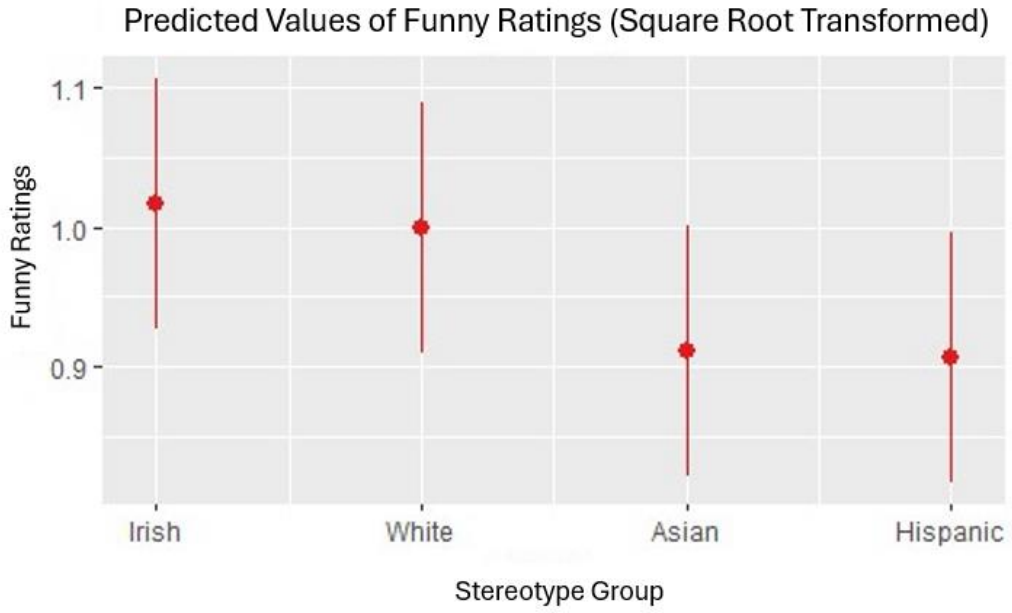


Figure 13.

Effect of Stereotype Group on Funny Ratings.



The effect of group in this model predicting funny ratings followed a similar pattern to the one found in the subversion model (see Figure 13). Of note, this was a very small effect, and a substantially higher amount of variability in responses (34%) was attributed to within-participant individual differences.

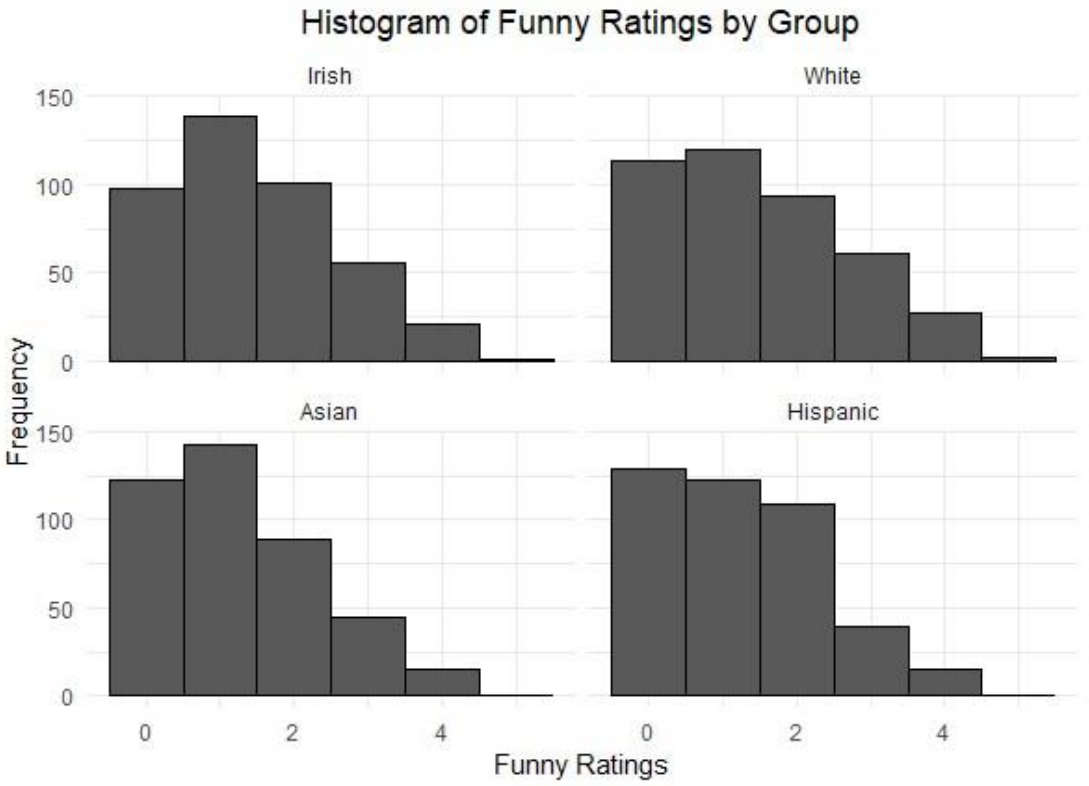
Exploratory Moderation Analyses (Funniness Ratings)

Additional exploratory analyses were conducted to investigate the moderating effect of group stereotype on meme funny ratings; an additional model predicting funny ratings was run including interaction terms between target group and both explicitness and distortion, as well as their interaction. This model showed significantly better fit than the model only including group as a predictor ($\chi^2 = 59.56$, $p < .001$). In this model, a very weak effect of explicitness emerged ($\beta = 0.24$, $p = .028$). Results showed that White memes were significantly funnier than Irish memes ($\beta = 0.37$, $p = .001$), Asian memes were significantly less funny than Irish memes ($\beta = -0.25$, $p = .024$) and Hispanic memes were not significantly different in funny ratings compared to Irish memes ($\beta = 0.03$, $p > .05$). Funny ratings by group are displayed in Figure 14.

Two groups showed moderation effects on explicitness when predicting funny ratings (see Figure 15) . For White memes, implicit memes were significantly funnier than White explicit memes ($\beta = -0.68$, $p < .001$). A similar pattern emerged for the Hispanic memes; implicit Hispanic memes were rated significantly funnier compared to explicit Hispanic memes ($\beta = -0.60$, $p < .001$). While the moderation effect of group on explicitness did not reach significance for the Asian memes ($\beta = 0.14$, $p > .05$), a reversed pattern emerged; explicit Asian memes were rated funnier than implicit Asian memes. Finally, a significant interaction emerged for Hispanic memes such that all memes were

Figure 14.

Histogram of Meme Funny Ratings by Group.

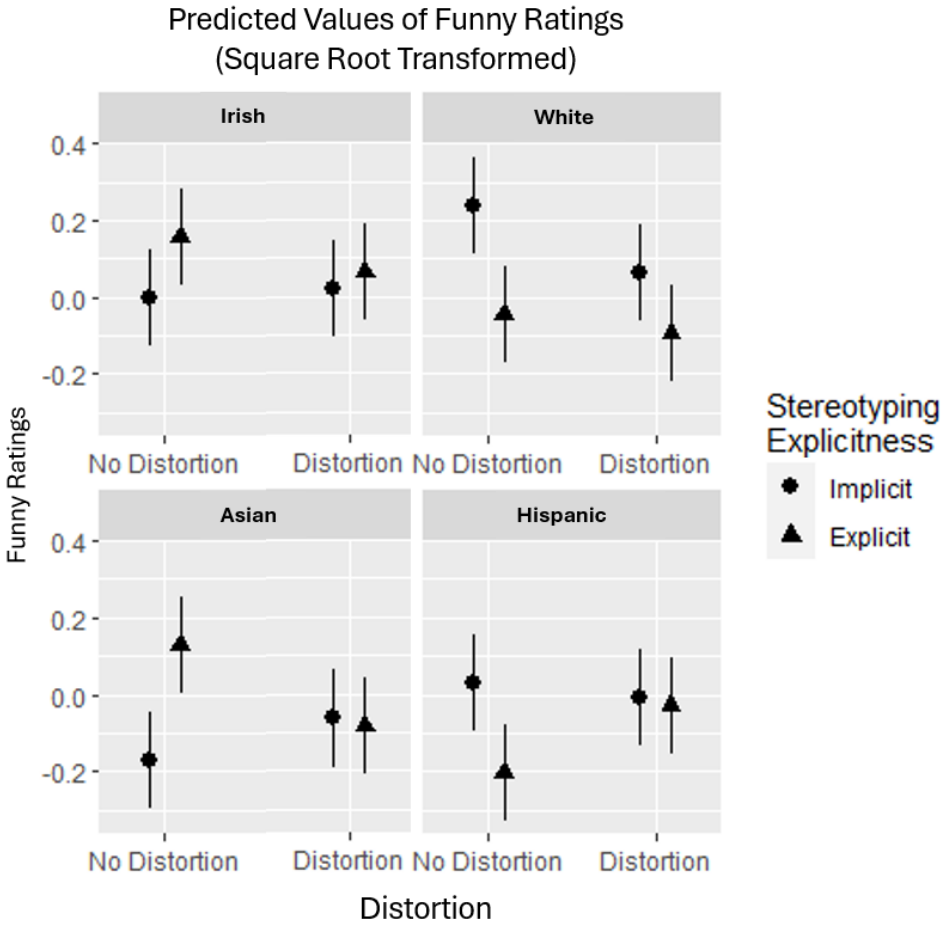


rated fairly similarly, except for non-distorted explicit memes which were rated significantly less funny ($\beta = 0.51$, $p = .022$).

Next, analyses were run separately for each target group and an additional predictor indicating whether participants reported being in-group or outgroup with respect to the target group was included in the models predicting funny ratings. Distortion, explicitness, and their interaction were the primary predictors as in previous models. In-group status of participants was also entered into the models as a predictor, as well as interaction terms between in-group status and the previous terms. Table 7 summarizes effects for analyses by group.

Figure 15.

Interaction Plot of Predicted Funny Ratings by Group.



For the Irish memes, none of the variables in the model were significant predictors of funny ratings; explicitness showed a marginally significant effect. For the White memes, results showed a significant interaction between explicitness and in-group status (β = -0.48, p < .05); for in-group members in particular, implicit White memes were rated funnier than explicit White memes and there was no real difference in ratings for outgroup members (see Figure 16). For Hispanic memes, results showed a similar significant effect of explicitness on funny ratings (β = -0.31, p < .05); implicit Hispanic memes were rated funnier than explicit Hispanic memes (see Figure 17, panel A).

Table 7.

Multilevel Regression Models Predicting Funny Ratings by Group

Funny Ratings (Square Root Transformed)				
	Irish	**White**	**Asian**	**Hispanic**
Predictors	*Std. Beta*	*Std. Beta*	*Std. Beta*	*Std. Beta*
Distortion	0.01	-0.26	0.18^{+}	0.03
Explicitness	0.14^{+}	-0.11	**0.40*** **	**-0.31***
In-Group	0.03	0.21	**0.50***	0.36^{+}
Distortion× Explicitness	-0.09	0.06	**-0.46****	0.32^{+}
Distortion × In-Group	0.14	-0.01	-0.13	-0.33
Explicitness × In-Group	0.20	**-0.48***	0.32	-0.14
Distortion × Explicitness × In-Group	-0.25	0.19	-0.13	0.02
Random Effects				
σ^2	0.25	0.27	0.20	0.23
$\tau_{00\ ID}$	0.15	0.18	0.19	0.20
ICC	0.38	0.40	0.49	0.46
$N_{\ ID}$	104	104	104	104
Observations	416	416	416	416
Marginal R^2/ Conditional R^2	0.015 / 0.387	0.045 / 0.426	0.078 / 0.527	0.028 / 0.477
AIC	769.45	806.31	717.37	765.56

Note. $^{+}$ indicates $p < .10$.* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Figure 16.

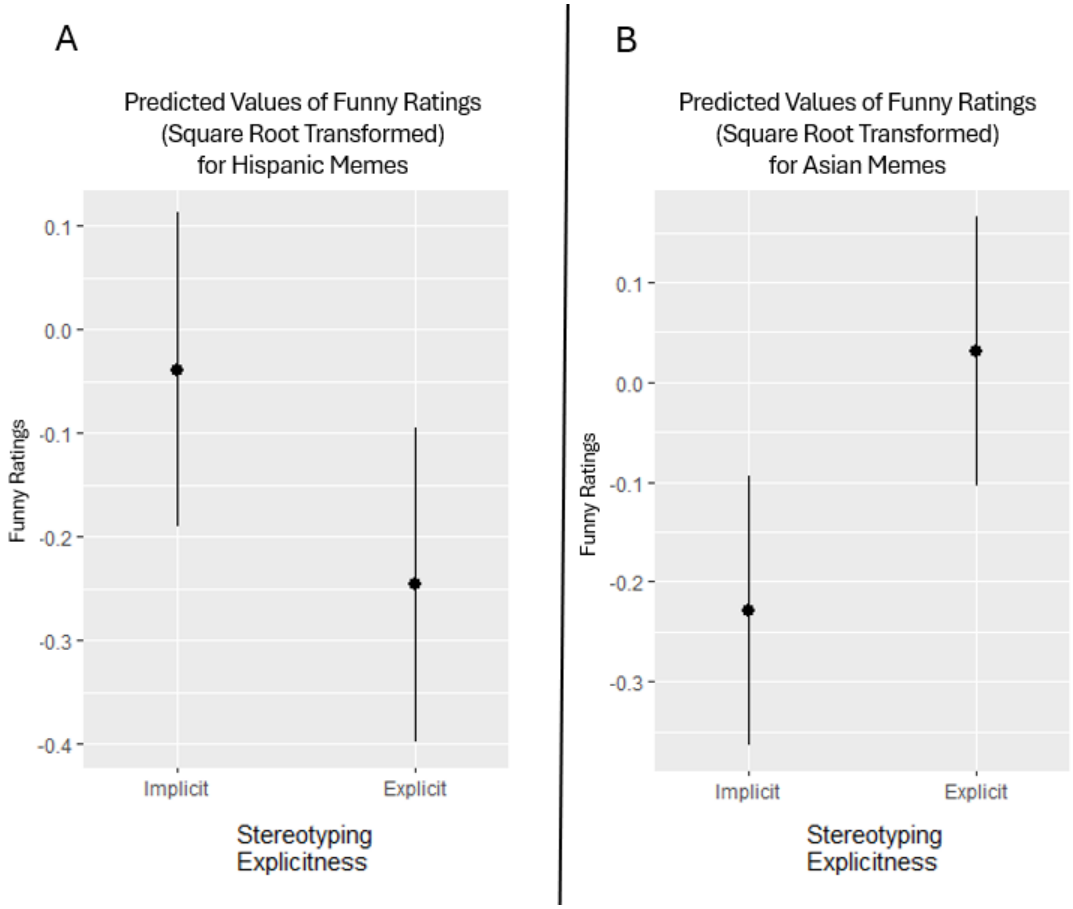
Interaction Plot of Predicted Funny Ratings for White Memes.

There was also a marginally significant effect of in-group status and a marginally significant interaction between distortion and explicitness for Hispanic memes, but these effects may not hold with a sufficiently powered sample.

Analyses for Asian memes showed several significant effects. As with Hispanic memes, Asian memes showed a significant effect of explicitness, however, it was in the opposite direction (β = .40, p < .001); explicit Asian memes were rated significantly funnier than implicit Asian memes (see Figure 17, panel B). In-group status showed a

Figure 17.

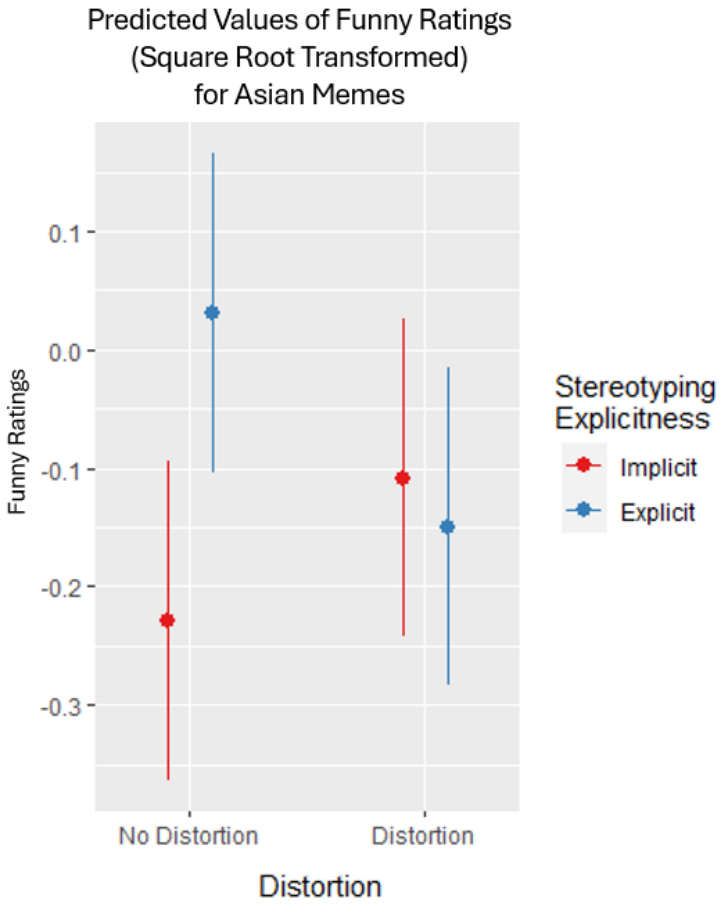
Effect of Stereotype Explicitness on Funny Ratings for Hispanic Memes (Panel A) and Asian Memes (Panel B)



significant positive effect on funny ratings of Asian memes (β = .50, p < .05); while there was a much smaller proportion of in-group Asian participants (>20%), these individuals rated Asian memes significantly funnier than outgroup individuals. Results also showed a significant interaction between distortion and explicitness for Asian memes, such that, explicitness was consequential only for non-distorted memes. Explicit, non-distorted Asian memes were rated significantly funnier than implicit, non-distorted Asian memes (see Figure 18). Finally, distortion showed a marginally significant effect for Asian memes (distorted memes funnier than non-distorted memes).

Figure 18.

Interaction Plot of Predicted Funny Ratings for Asian Memes.



Results of Model Predicting Offense Ratings

The final section of pre-registered analyses concerns meme offense ratings. As mentioned previously, offense ratings were highly positively skewed and severely zero inflated. Therefore, a series of multilevel negative binomial regression models were fitted to account for over-dispersion. The model predicting offense ratings from distortion, explicitness, their interaction, and participant ID (random intercept) did not perform significantly different than the base model with only participant ID entered as a random

factor. Including the group variable into the model with all predictors entered resulted in a significantly better fitted model compared to the base model ($\chi^2 = 34.8$, $p < .001$). Removing the interaction term from this model led to a better fit, which was not significantly different from the complete model with all terms included. Therefore, results are presented in Table 8 for the simpler model including only distortion, explicitness, and group as fixed predictors. Because White memes were rated least offensive, White now serves as the reference group for the following analyses.

Table 8.

Negative Binomial Multilevel Regression Predicting Offense Ratings.

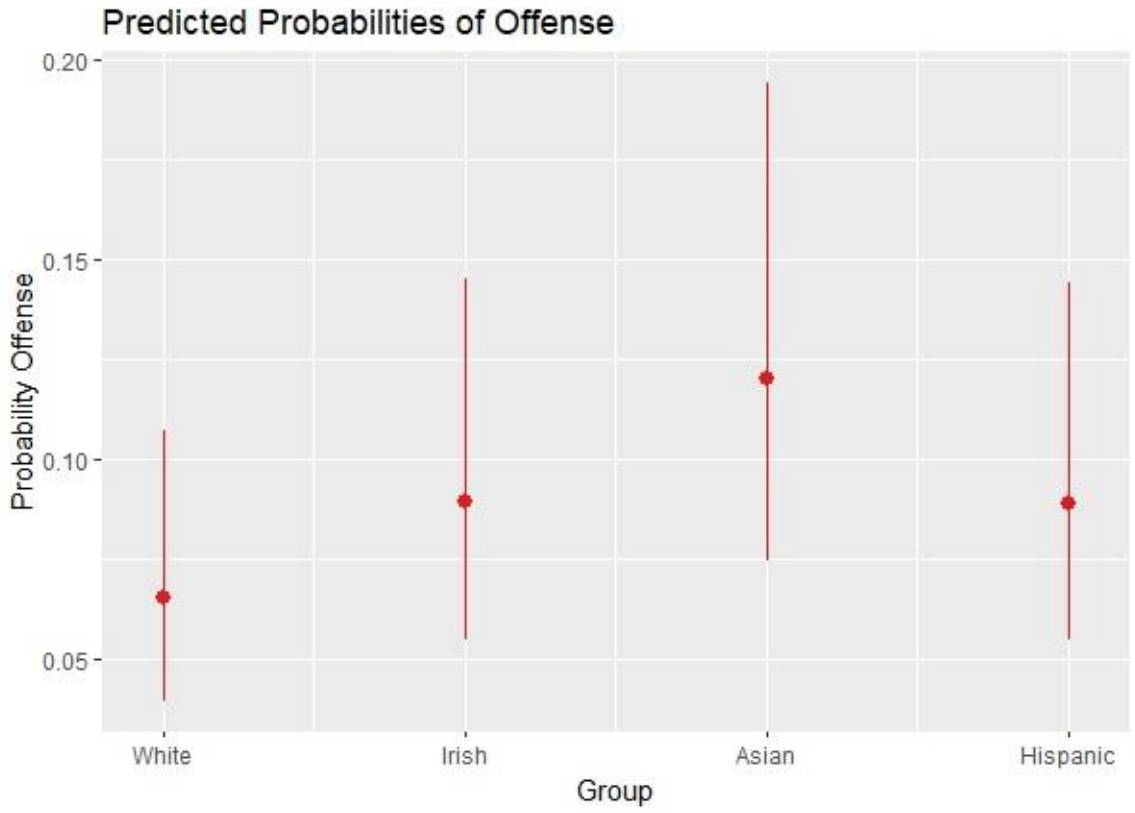
	Offense Ratings			
Predictors	*Incidence Rate Ratios*	*std. Beta*	*standardized CI*	*p*
(Intercept)	0.07	0.07	0.04 – 0.11	**<0.001**
Distortion	1.22	1.22	1.04 – 1.43	**0.015**
Explicitness	1.05	1.05	0.89 – 1.23	0.569
Group [Irish]	1.37	1.37	1.07 – 1.76	**0.013**
Group [Asian]	1.84	1.84	1.46 – 2.33	**<0.001**
Group [Hispanic]	1.36	1.36	1.06 – 1.74	**0.015**
Random Effects				
σ^2	2.37			
$\tau_{00\ ID}$	3.25			
ICC	0.58			
$N_{\ ID}$	104			
Observations	1664			
Marginal R^2 / Conditional R^2	0.010 / 0.583			
AIC	1987.330			

Note. White is the reference group.

Results did not show any support for Hypothesis 3a (interaction). In terms of Hypothesis 3b, contrary to what was expected, there was a positive effect of distortion such that distortion positively predicted offense ratings. Finally, no support was found for Hypothesis 3c; where I expected explicit jokes to be rated significantly more offensive, explicitness was not predictive of offense ratings. A majority of the variance in offense ratings (58%) was attributed to intra-participant individual differences. As with the other outcomes, the group variable was found to significantly predict offense ratings. In the case of offense ratings, Irish, Hispanic, and Asian memes were all rated significantly more offensive than White memes (see Figure 19).

Figure 19.

Effect of Stereotype Group on Offense Ratings.



Exploratory Moderation Analyses (Offensive Ratings)

Additional exploratory analyses were conducted to investigate the moderating effect of group stereotype on meme offense ratings; an additional model predicting offense ratings was run including interaction terms between target group and both explicitness and distortion, as well as their interaction. The model looking at moderation did not display significantly different fit compared to the model including group as an additional predictor ($\chi^2 = 16.50$, $p > .05$). In this moderation model, distortion was no longer a significant predictor. Asian memes were the only group displaying significantly more offensive ratings compared to White memes ($\beta = 1.61$, $p = .016$). The only significant moderation by group that emerged was on the effect of distortion in the case of Irish memes (see Figure 20). Distorted Irish memes were more likely to result in higher offense ratings than non-distorted Irish memes ($\beta = 2.06$, $p = .011$). Offense ratings by group are displayed in Figure 21.

Figure 20.

Moderation of Distortion by Group on Offense Ratings

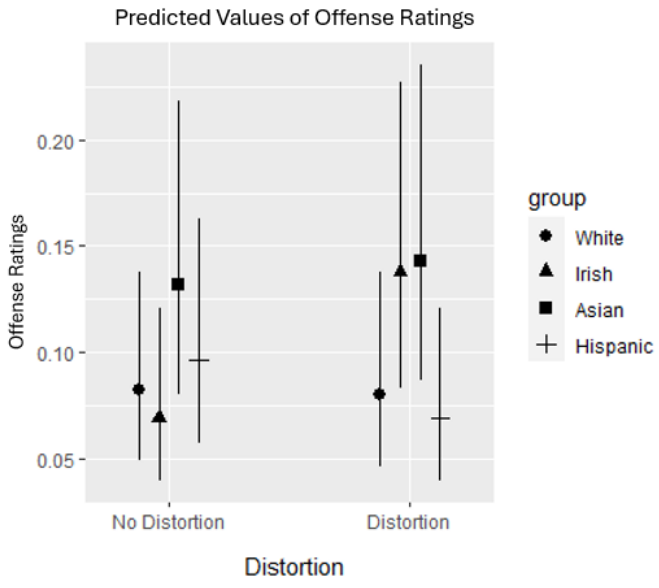
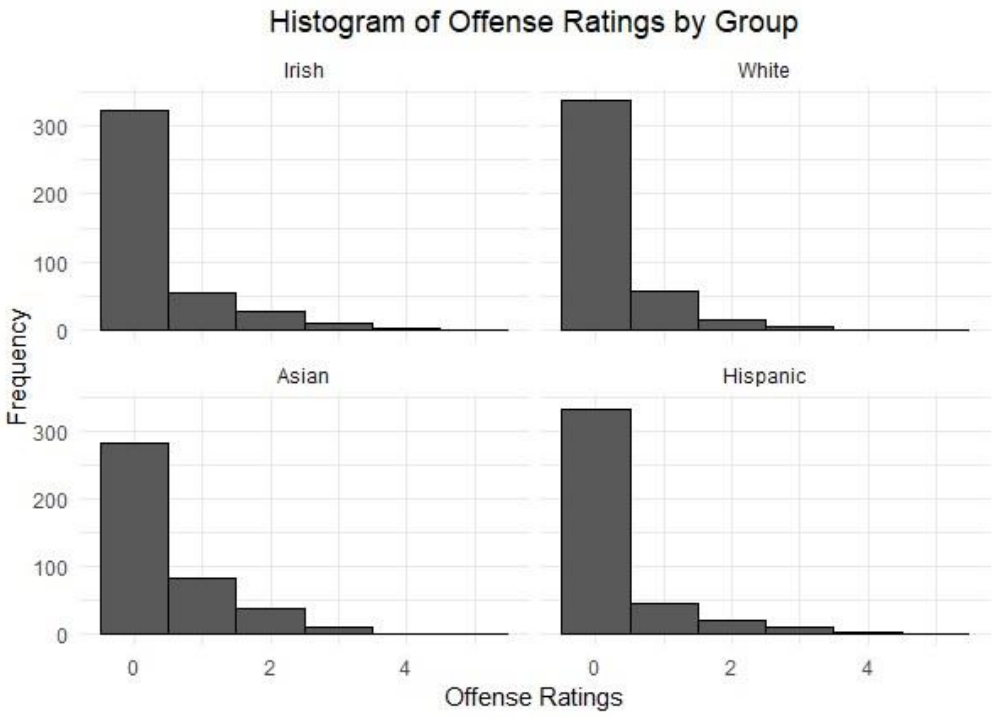


Figure 21.

Histogram of Meme Offensiveness Ratings by Group.



Next, analyses were run separately for each target group and an additional predictor indicating whether participants reported being in-group or outgroup with respect to the target group was included in the models predicting offense ratings. Distortion and explicitness were the primary predictors as in previous models predicting offense. In-group status of participants was also entered into the models as a predictor, as well as interaction terms between in-group status and the previous terms. Table 9 summarizes effects for analyses by group.

For the White and Asian memes, none of the variables in the model were significant predictors of offense ratings. For the Irish memes, results showed a significant effect of distortion on offense ratings ($\beta = 1.84$, $p < .01$), such that distorted Irish memes

were rated more offensive compared to non-distorted Irish memes. For Hispanic memes, results showed a significant effect of explicitness on offense ratings ($\beta = 1.54$, $p < .05$); explicit Hispanic memes were rated more offensive than implicit Hispanic memes.

Table 9.

Negative Binomial Multilevel Regression Models Predicting Offense Ratings by Group.

	Offense Ratings			
	Irish	**White**	**Asian**	**Hispanic**
Predictors	*Std. Beta*	*Std. Beta*	*Std. Beta*	*Std. Beta*
Distortion	**1.84****	1.00	1.14	1.06
Explicitness	0.93	1.36	1.06	**1.54***
In-Group	3.50^{+}	2.14	1.08	0.53
Distortion × In-Group	0.81	1.22	0.98	0.47
Explicitness × In-Group	1.19	0.52^{+}	0.67	0.57
Random Effects				
σ^2	2.63	3.10	2.00	2.74
$\tau_{00\ ID}$	3.52	4.23	2.80	3.50
ICC	0.57	0.58	0.58	0.56
N_{ID}	104	104	104	104
Observations	416	416	416	416
Marginal R^2/ Conditional R^2	0.04 / 0.589	0.013 / 0.583	0.003 / 0.585	0.063 / 0.589
AIC	514.94	437.04	641.06	511.09

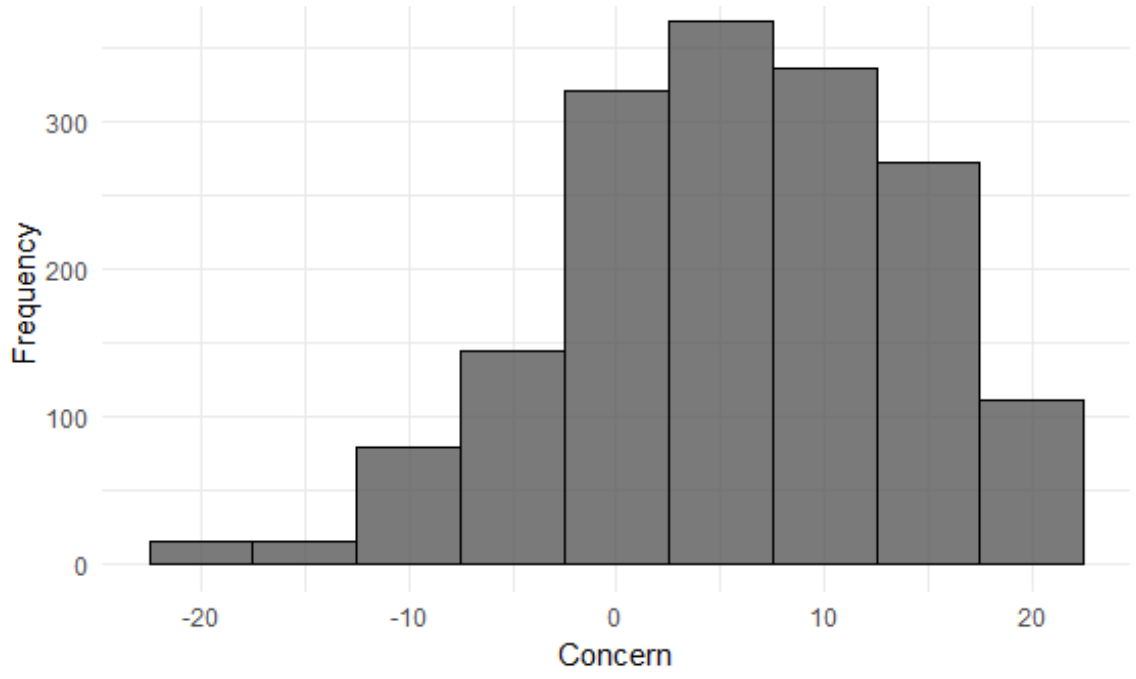
Note. $^{+}$ indicates $p < .10$.* indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Results of Additional Exploratory Analyses

A few additional analyses were conducted to explore the effect of concern over appearing prejudiced on the three key dependent variables of this study (subtext judgments, meme funniness, and meme offensiveness). Distribution of concern scores appear in Figure 22 below. Additionally, models examined the relationship between the three outcomes, controlling for previous predictors (explicitness, distortion, group), and accounting for nesting of responses at the participant level: subversion predicted by funny and offense ratings, funny ratings predicted by subversion and offense ratings, and offense ratings predicted by subversion and funny ratings. Moderation by group was also investigated in these models.

Figure 22.

Histogram of Concern over Appearing Prejudiced.



Concern did not significantly predict subversion ratings. It did, however, interact significantly with Hispanic memes ($\beta = 1.54$, $p < .05$). Higher scores on concern were associated with greater probability of subversion only for Hispanic memes; the pattern was reversed for White and Irish memes, and there was no difference for Asian memes. A follow up analysis of only Hispanic memes showed no significant effect of concern. Funny and offense ratings did not significantly predict subversion ratings, nor did they significantly interact with target group in predicting subversion.

Concern also did not significantly predict funny ratings. Offense ratings did emerge as a significant predictor of funny ratings ($\beta = -0.28$, $p < .001$), with these variables being inversely related. Offense significantly interacted with the White memes ($\beta = 0.21$, $p = .001$). Funny ratings for Asian, Irish, and Hispanic memes showed the expected pattern in which lower offense scores predicted higher funny ratings, but this pattern was nonexistent for White memes (no difference between high and low offense). Hispanic memes showed a significant interaction with subversion ratings when predicting funny ratings ($\beta = 0.24$, $p = .048$). These memes were the only ones for which subversion (as opposed to endorsement) ratings were associated with higher funny ratings.

Finally, concern did significantly predict offense ratings ($\beta = 2.76$, $p < .001$). Higher concern scores were associated with higher offense ratings. Concern also significantly interacted with Asian memes in predicting offense ($\beta = .68$, $p = .007$). For Irish, White, and Hispanic memes, both concern scores one standard deviation below the mean and mean concern scores were associated with lower offense ratings compared to Asian memes that displayed much higher offense ratings even at low levels of concern. As in the previous model, funny ratings were significantly associated with offense

ratings; higher funny ratings predicted lower offense ratings (β = 0.79, p = .009). Funny ratings also significantly interacted with Hispanic memes in predicting offense (β = .76, p = .031). For this group in particular the effect of funny ratings on offense was much more prominent than the other three groups, which each displayed greater overlap in offense scores across low to high funny ratings (no real difference in offense between funny scores one standard deviation below and above the mean).

CHAPTER 5

DISCUSSION

The aim of this study was to evaluate the effects of two facets of stereotype humor, explicitness of stereotyping and stereotype distortion, on judgments of stereotype subtext (endorsement/subversion) in memes about four different groups (Asian, Hispanic, Irish, White) and associated group stereotypes. Additionally, I explored how these factors influenced ratings of meme funniness and offensiveness, as endorsement and subversion judgments are theoretically implicated in the emotional responses (amusement, offense) elicited by stereotype humor ("sword" vs. "shield" metaphor; Rappaport, 2005).

I hypothesized that explicitness and distortion would significantly interact resulting in greater ratings of subversion, higher funny ratings, and lower offense ratings only for explicitly distorted memes. These memes directly present viewers with stereotype information and challenge it head-on through the presentation of an alternate, exaggerated, and ridiculous instantiation of the stereotype. Non-distorted memes, where the stereotype is presented at face value, were expected to lead primarily to judgments of endorsement regardless of level of explicitness because they require viewers to coactivate schematic information surrounding a group with their group stereotype, thus reinforcing the association in the viewer's mind. Likewise, implicit distorted memes require a viewer to self-generate relevant stereotype schemas in order to comprehend the joke; presumably this would result in judgments of endorsement, based on the self-generated stereotype thought, even though distorting a stereotype is a means of challenging the typical schematic representation through a humorous alternate framing of the stereotype.

Funny ratings were expected to track subversion judgments, since joking about stereotypes is taboo, unless the joke directly confronts the stereotype (Kramer, 2020). Explicitness was hypothesized to only have an impact on offensiveness ratings, with explicit memes leading to greater offense ratings, since outright expressions of stereotyping are typically condemned in this day and age as prejudicial (Devine et. al., 2002; Neuberg et al., 2020).

Subversion Ratings.

Results demonstrated that stereotype distortion (as opposed to no distortion) and, contrary to hypotheses, explicitness of stereotyping (as opposed to implicitness) both significantly positively predicted ratings of subversion across the different stimuli; I did not find evidence of a significant interaction. Thus, in terms of subtext judgments, Hypothesis H1a (significant interaction) and H1c (no effect of explicitness) were rejected, and results only supported Hypothesis H1b (significant effect of distortion). Although I expected target group not to make a difference in how people process stereotype humor, because all stereotypes are schematic representations activated similarly through the process of spreading activation, a main effect of group emerged, such that memes that targeted the two more marginalized groups (Asian, Hispanic) investigated in this study were far less likely to be interpreted as subversive compared to the more privileged groups (White, Irish). This pattern might be attributed to the fact that two-thirds of the sample identified as White. Members of majority groups commonly experience concern over appearing biased to others (Borgella et al, 2020; Plant & Devine, 1998). This concern may have translated to White participants being over cautious in

their interpretations of humor about Asian and Hispanic stereotypes, defaulting to more straightforward judgments of endorsement (and relatedly, greater offense).

More broadly, previous research by Jeffries and colleagues (2012) provides cross-cultural evidence that individuals are inclined to perceive criticisms of higher status groups as more permissible than criticisms of lower status groups ("David and Goliath" principle). Joking about negative stereotypes, in particular, can be interpreted as a form of criticism, since it links the group to some negative stereotyped information. This difference in normative permissibility has the potential to interrupt the humor process for jokes targeting lower status groups, making it harder for people to get into a humorous mindset when encountering this type of humor (McGraw & Warren, 2010), and more difficult to cognitively shift to whatever alternative conceptual framework is established by the joke (i.e., too much dissonance). This, ironically, could result in negative consequences in the context of stereotype humor intended to subvert and challenge stereotypes. In one case, minority groups joking about their ingroup may find it more difficult to successfully communicate subversion of their group's stereotypes, leading to offense in audiences. In another case, higher status jokers attempting to critique and challenge stereotypes about outgroup minority members could be perceived as biased or prejudicial for endorsing stereotypes, even though their intent was to subvert.

A more nuanced pattern of results emerged in subsequent analyses that split up and examined subtext judgments by target group. Results of these more granular analyses demonstrated that the primary drivers of the significant effects of explicitness and distortion on subversion were responses to the memes targeting Asian and Hispanic stereotypes. Variations in presentation of the stereotype did not have any impact on

interpretations of memes targeting the White stereotype. Distortion did remain a significant predictor of ratings of subversion for Irish memes, in the expected direction, but to a much lower degree than for the marginalized target groups. Why did distortion and explicitness only impact subtext interpretations of the marginalized groups' memes? One possibility is that these factors are only relevant in contexts where stereotype jokes are targeting outgroup members. Individuals are biased to view members of outgroups as more homogeneous (less diverse) compared to one's own group (Anthony et al., 1992). Since the sample was primarily White, it may be that it was easier to discount the White stereotype as ridiculous (subversive)—or not applicable to the group as a whole—regardless of stereotype explicitness or distortion compared to the memes about the other minoritized group stereotypes. Even though analyses did not reveal significant effects of in-group status, the sample was not sufficiently balanced across groups to fully investigate this variable.

Another possibility is that belief in the validity of the stereotype, which we did not measure in this study, could alter how individuals process subtext in stereotype humor. The stereotypes about the more privileged groups (drink a lot, can't dance) might have been interpreted as more ridiculous to begin with, thus nullifying any influence of stereotype presentation, whereas the stereotypes about the marginalized groups (demanding parents, parental violence) may be viewed as possessing some "kernel of truth" (Jussim et al., 2015) and therefore are more deeply considered or processed when judging subtext. Although you do not need to believe in a stereotype to comprehend a stereotype joke, level of belief may lead to differences in capacity to interpret underlying

messaging of the joke. It is easier to challenge beliefs you do not hold strongly than beliefs you are more strongly invested in (Klayman, 1995).

Thinking back to the expanded stereotype humor process model, results of this study support the notion that some stereotypes, particularly those about minoritized groups (Asian, Hispanic) are perceived as being more societally, culturally, or normatively taboo to joke about. Presumably, some portion of participants felt dissonance (i.e., "can't joke about this group") when engaging with the Asian and Hispanic stereotype memes leading to greater ratings of offense for these memes. Those that did not feel a significant level of dissonance, either because they were a part of the group or just less concerned with appearing prejudiced, then relied on the inferred messaging or subtext of the meme driven by heuristic processing related to distortion of the stereotype and explicitness of stereotyping.

Funny and offense ratings did not significantly relate to subversion ratings, nor did they significantly interact with target group in predicting subversion (except for ratings of subversion of Hispanic memes leading to higher funny ratings). This finding is counter to the theorized link between subtext and resulting emotional reactions to stereotype humor proposed in the "sword and shield" metaphor of stereotype humor (Rappaport, 2005). Humor that is interpreted as a "shield" for protecting against prejudice (subversive humor) should be viewed more positively, and lead to greater amusement, than humor which act as a "sword" by weaponizing stereotypes against groups (endorsing/disparaging humor), which should lead to greater offense. However, the current study along with previously reviewed work (Miller et al., 2019; Saucier et al., 2018) point to greater complexity in processing stereotype humor; it is not so black and

white for people and this ambiguity most likely results in a more complex suite of emotional responses for any stereotype humor which exists in the gray area between clearly funny and clearly offensive. Stimuli in this study were designed precisely to exist in this gray area, and therefore reflect this separation between judgments of subtext and resulting emotional responses. Additionally, participants were forced to make an endorsement or subversion judgment before rating memes for funniness and offensiveness, but in terms of the stereotype humor process model, participants that were more highly concerned and felt more dissonance may have felt stronger levels of offense at the time of making the endorsement or subversion rating, before even deciding on subtext. For this group of participants, their rating of subtext would have been independent of offense, such that even those that saw the meme as subversive would have rated it more highly offensive.

Funny Ratings.

Primary analyses predicting funny ratings only showed a significant effect of target group, leading to the rejection of Hypotheses 2a (significant interaction) and 2b (significant effect of distortion), but supporting 2c (no effect of explicitness). However, splitting analyses by group again showed differences in patterns of the effects of explicitness and distortion. Neither factor mattered for Irish or White memes. Conversely, explicitness was a significant predictor of funny ratings for both Hispanic and Asian memes, but in opposite directions. Explicit Asian memes were rated funnier than implicit Asian memes, and implicit Hispanic memes were rated funnier than explicit Hispanic memes. Interestingly, White participants rating White memes (in-group humor) showed a

similar pattern of results as Hispanic memes, in that, implicit White memes were rated funnier than explicit White memes.

What might account for this pattern? Xing (2019) found that jokes about negative stereotypes were more offensive but also funnier than jokes about positive stereotypes. Our sample reported the Asian stereotype as the most offensive and also the most familiar. Since it was the most familiar stereotype, the ease with which participants accessed the Asian stereotype information in the explicit conditions should have contributed to a faster "aha moment" for participants encountering these memes to get the joke, resulting in funnier ratings. Even though the joke was about a more offensive stereotype (according to the sample), this might have just amplified funny ratings in a sort of paradoxical self-protective line of reasoning (i.e., "this is offensive, but I fluidly understood it, and I'm not prejudiced, so it must be really funny").

The fact that implicit White and implicit Hispanic memes were rated funnier than explicit ones might actually be reflective of the same process; these two groups were the highest represented groups in the sample (65% and 30% respectively) and might have had greater ease accessing the stereotypes about their groups. For these groups in particular, implicit stereotype memes would have reflected group-level encrypted or "inside jokes" which are perceived as signals that the joker and listener share a similar sense of worldview (Flamson & Barrett, 2008). This presumed similarity, even when the joker is not directly identified, is likely to increase the permissibility of laughing at a stereotype joke. Thai and colleagues (2019) find evidence for this in their finding that it is more socially acceptable for people to joke about their own group.

Offense Ratings.

Initial analyses for offense demonstrated a significant effect of distortion, but in a direction contrary to hypotheses (H3b): distorted memes were rated more offensive than non-distorted memes. There was no significant interaction (H3a) between explicitness and distortion, no effect of explicitness (H3c), and therefore, none of the offense hypotheses held up. The effect of distortion appearing in the opposite direction from what was predicted could reflect high concern over appearing prejudiced in this sample. This is supported by the fact that concern emerged as a significant predictor of offense in exploratory models. The sample's scores on the MCPR concern subscale were negatively skewed and relatively high. The college sample was also very young and presumably highly liberal. Distorting stereotypes might have been interpreted as permitting the norm that prejudiced beliefs are allowed to be poked fun at, played with, or just not that big of a deal (Ford & Ferguson, 2004), which may have just been too much of a violation (McGraw & Warren, 2010) for this sample that has grown up in an ever increasing politically correct social environment.

Group also emerged as a main predictor of offense ratings, but this time Hispanic, Asian, and Irish memes were all significantly different (more offensive) compared to White memes. This may once again come down to the sample being primarily White and rating jokes about their in-group as less offensive compared to the other groups, including the sub-group Irish, which was only represented by 12.5% of the participants. When analyses were split by group, the only two significant effects were distortion in the case of Irish memes (distorted memes were more offensive than non-distorted memes) and explicitness in the case of Hispanic memes (explicit memes were more offensive than

implicit memes). It is unclear why these effects occurred for these groups in particular, as opposed to following similar minoritized vs. privileged group patterns as with previous outcomes. It may come down to the two stereotypes in question (drink a lot, parental violence) being more objectively negative as compared to the other stereotypes (demanding parents, can't dance). Distorting the Irish meme amplified the severity of drunkenness, while explicitly stating that Hispanic mom's hit their kids is far more severe than implying it. This could have pushed participants to process this humor as more offensive rather than amusing.

Strengths and Limitations

This preregistered study was the first attempt to rigorously investigate the influence of objective features of stereotype humor on interpretations of the underlying messaging of the joke, as well as emotional reactions. Moreover, it was done in the context of less disparaging, less offensive, and less overtly prejudiced stereotypes that were well known to participants, statistically prevalent in the target groups, and documented in research. This was purposefully done to reflect the type of humor, which is more frequently encountered in day-to-day life, as well as to avoid exposing participants to stereotype humor that relies on harmful and obscene racial slurs. The design allowed for comparison between privileged versus marginalized groups and did not rely on researcher-selected subversive humor, which may not have been interpreted the same way by all participants; instead, several memes were systematically manipulated on the constructs of interest, allowing for a well-powered, within-subjects investigation of the processes at play when individuals encounter stereotype humor.

There were, of course, some limitations. While memes did fall into a range of funniness, this range could have been more normally distributed. Plus, memes are already outdated as a form of humor, especially in this college age sample, and future work will need to rely on stimuli that participants encounter more frequently that is more relevant to their age group (e.g., tiktoks, IG reels), or explore these research questions in a more diverse sample. The sample was primarily white, liberal, and young; would results hold in more general audiences? Similarly, while I achieved memes that were low in offense, as was the goal, the distribution was severely skewed with a very high rate of "not at all offensive" ratings, making it more difficult to explore the relation between this variable and the other outcomes of interest. The sample was also insufficiently distributed among identification with the target group, leading to underpowered exploration of moderation. Additionally, the design may have resulted in order and fatigue effects since memes were presented in rapid succession and subtext judgements and emotion ratings were not counterbalanced. Future work could implement more realistic contexts (e.g., scrolling on a feed with a variety of posts around the humor) or measure daily exposure to stereotype humor in a more naturalistic way (e.g., EMA).

Future Directions

This project serves as the foundation for a fruitful research line with many possible future avenues to explore. The first is including information about the joker's identity, since this variable has previously been shown to alter perceptions of stereotype humor (Thai et al., 2019; Langley & Shiota, in prep). Research has repeatedly demonstrated that stereotype humor is rated more positively when it is coming from an in-group member. Results of the current study suggest that individuals evaluate humor

about minority groups as less likely to be subversive, less funny, and more offensive. I predict that, especially for jokes about minority stereotypes, knowing the joke is coming from an in-group member will reverse these patterns and jokes will be rated more subversive, funnier, and less offensive because in-group jokers are permitted to comment on their own group and will be less likely to be inferred as actually intending to support a negative stereotype about their own group (i.e., presumed sarcasm). In terms of the stereotype humor process model, joker identity would influence both dissonance related to concepts presented in the stereotype (i.e., lower dissonance for in-group jokers) and how personally taboo or concerning the punchline is evaluated to be (i.e., in-group jokers more subversive, less concerning).

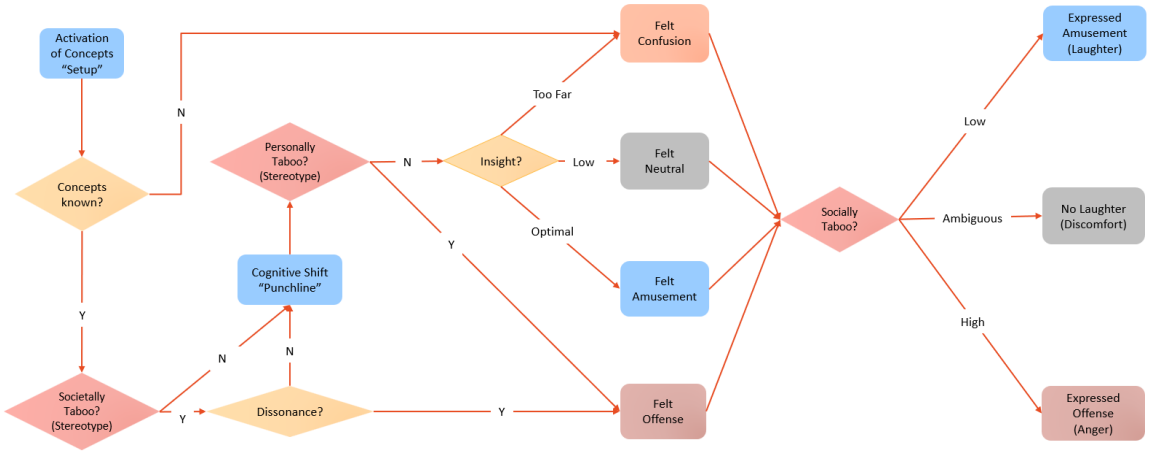
Another future direction involves investigating reactions to stereotype humor using dynamic assessments of videos of standup routines through a rating dial that would allow for collection of time series data, allowing for a finer grain exploration of the processes at play, yet again in a well-powered within-subjects design. This would also facilitate investigating identities beyond nationality, race, and ethnicity since there are a plethora of videos of standup comedians discussing any and every type of identity available online. Future work examining whether effects hold in the case of jokes about "positive" stereotypes; effects of stereotype humor that distorts stereotypes through other means including negation, minimizing, or presentation of counter evidence; and how individuals with intersectional backgrounds process this type of humor is also needed as these are huge holes in the literature on stereotype humor.

Finally, beyond intrapersonal reactions to stereotype humor, more work is needed on interpersonal reactions surrounding how people react to, combat, or appreciate real

encounters of stereotype humor ("in the wild") along with the downstream social consequences of these behavioral responses in the ways of traditional social psychological research. Figure 23 below offers a complete model of stereotype humor, including the social influence of other individuals being present on expressed emotion. The initial expanded model focuses on internal cognitive processes that lead to differences in felt emotion, however, there are instances where felt emotion and expressed emotion are at odds. For example, someone might process a stereotype joke as amusing, but when their expression of amusement will depend on how taboo it would be to laugh in front of present company. If taboo is low, as is the case when everyone else is laughing or the group is perceived as one that accepts the stereotype joke, laughter is likely. However, if social taboo is more ambiguous (e.g., told in mixed company), laughter may be stifled and suppression of amusement could lead to discomfort. If taboo is high, felt amusement could transform into displayed offense in order to remain accepted by the group. Similar processes apply to felt offense, confusion, and neutrality.

Figure 23.

Complete Stereotype Humor Process Model



REFERENCES

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22*(3), 261-295.

Anthony T., Copper C., Mullen B. (1992). Cross-racial facial identification: A social cognitive integration. *Personality and Social Psychology Bulletin*, 18, 296–301.

Apte, M. L. (1985). *Humor and laughter: An anthropological approach*. Ithaca, NY: Cornell University Press.

Attardo, S. (1997). The semantic foundations of cognitive theories of humor. *Humor, 10*(4), 1997, pp. 395-420.

Attardo, S., & Raskin, V. (1991). Script theory revis (it) ed: Joke similarity and joke representation model.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230.

Bates, D. M. (2010). lme4: Mixed-effects modeling with R.

Baumgartner, J. C., & Morris, J. S. (2008). One "nation," under Stephen? The effects of the Colbert Report on American youth. *Journal of Broadcasting & Electronic Media, 52*(4), 622-643.

Beeman, W. O. (1999). Humor. *Journal of Linguistic Anthropology, 9*(1/2), 103-106.

Block, J. (1982). Assimilation, accommodation, and the dynamics of personality development. *Child Development*, 281-295.

Borgella, A. M., Howard, S., & Maddox, K. B. (2020). Cracking wise to break the ice: The potential for racial humor to ease interracial anxiety. *Humor, 33*(1), 105-135.

Braun, A., & Preiser, S. (2013). The impact of disparaging humor content on the funniness of political jokes. *Humor, 26*(2), 249-275.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407.

Denham, S. L., Farkas, D., Van Ee, R., Taranu, M., Kocsis, Z., Wimmer, M., ... & Winkler, I. (2018). Similar but separate systems underlie perceptual bistability in vision and audition. *Scientific Reports, 8*(1), 1-10.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5.

Devine, P. G., Plant, E. A., Amodio, D. M., Harmon-Jones, E., & Vance, S. L. (2002). The regulation of explicit and implicit race bias: the role of motivations to respond without prejudice. *Journal of Personality and Social Psychology, 82*(5), 835.

Dovidio, J. F., & Gaertner, S. L. (2010). Intergroup bias. In S. T. Fiske,D. T. Gilbert, & G. Lindzey (Eds.),*Handbook of social psychology*(5thed., pp. 1084–1121). Hoboken, NJ: Wiley.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin, 23*(3), 316-326.

Dynel, M. (2009). Beyond a joke: Types of conversational humour. *Language and Linguistics Compass, 3*(5), 1284-1299.

Fine, G. A., & Soucey, M. D. (2005). *Joking cultures: Humor themes as social regulation in group life.*

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in Experimental Social Psychology (*Vol. 23, pp. 1-74). Academic Press.

Flamson, T., & Barrett, H. C. (2008). The encryption theory of humor: A knowledge-based mechanism of honest signaling. *Journal of Evolutionary Psychology, 6*(4), 261-281.

Ford, T. E., Boxer, C. F., Armstrong, J., & Edel, J. R. (2008). More than "just a joke": The prejudice-releasing function of sexist humor. *Personality and Social Psychology Bulletin, 34*(2), 159-170.

Ford, T. E., & Ferguson, M. A. (2004). Social consequences of disparagement humor: A prejudiced norm theory. *Personality and Social Psychology Review, 8*(1), 79-94.

Ford, T. E., Richardson, K., & Petit, W. E. (2015). Disparagement humor and prejudice: Contemporary theory and research. *Humor, 28*(2), 171-186.

Ford, T. E., Woodzicka, J. A., Triplett, S. R., Kochersberger, A. O., & Holden, C. J. (2014). Not all groups are equal: Differential vulnerability of social groups to the prejudice-releasing effects of disparagement humor. *Group Processes & Intergroup Relations, 17*(2), 178-199.

Fraley, B., & Aron, A. (2004). The effect of a shared humorous experience on closeness in initial encounters. *Personal Relationships, 11*(1), 61-78.

Gabora, L., & Kitto, K. (2017). Toward a quantum theory of humor. *Frontiers in Physics, 4*, 53.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78*(4), 708.

Gervais, M., & Wilson, D. S. (2005). The evolution and functions of laughter and humor: A synthetic approach. *The Quarterly Review of Biology, 80*(4), 395-430.

Hall, J. A. (2013). Humor in long-term romantic relationships: The association of general humor styles and relationship-specific functions with relationship satisfaction. *Western Journal of Communication, 77*(3), 272-292.

Hart, M. T. (2007). Humour and social protest: An introduction. *International Review of Social History, 52*(S15), 1-20.

Hastie, R. (1981). Schematic principles in human memory. In *Social cognition* (pp. 39-88). Routledge.

Hay, J. (2000). Functions of humor in the conversations of men and women. *Journal of Pragmatics, 32*(6), 709-742.

Holmes, J., & Marra, M. (2002). Having a laugh at work: How humour contributes to workplace culture. *Journal of Pragmatics, 34*(12), 1683-1710.

Jeffries, C. H., Hornsey, M. J., Sutton, R. M., Douglas, K. M., & Bain, P. G. (2012). The David and Goliath Principle: cultural, ideological, and attitudinal underpinnings of the normative protection of low-status groups from criticism. *Personality and Social Psychology Bulletin, 38*(8), 1053-1065.

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.

Kao, J. T., Levy, R., & Goodman, N. D. (2016). A computational model of linguistic humor in puns. *Cognitive Science, 40*(5), 1270-1285.

Keltner, D., Capps, L., Kring, A. M., Young, R. C., & Heerey, E. A. (2001). Just teasing: a conceptual analysis and empirical review. *Psychological Bulletin, 127*(2), 229.

Keltner, D., Young, R. C., Heerey, E. A., Oemig, C., & Monarch, N. D. (1998). Teasing in hierarchical and intimate relations. *Journal of Personality and Social Psychology, 75*(5), 1231.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation, 32*, 385-418.

Knight, N. K. (2013). Evaluating experience in funny ways: How friends bond through conversational hum. *Text & Talk, 33*(4-5), 553-574.

Kochersberger, A. O., Ford, T. E., Woodzicka, J. A., Romero-Sanchez, M., & Carretero-Dios, H. (2014). The role of identification with women as a determinant of amusement with sexist humor. *Humor, 27*(3), 441-460.

Koestler, A. (1964). *The act of creation: A study of the conscious and unconscious processes of humor, scientific discovery and art*.

Kramer, C. A. (2020). Subversive humor as art and the art of subversive humor. *The Philosophy of Humor Yearbook, 1*(1), 153-179.

Kulka, T. (2007). The incongruity of incongruity theories of humor. *Organon F, 14*(3), 320-333.

Kounios, J., & Beeman, M. (2009). The Aha! moment: The cognitive neuroscience of insight. *Current Directions in Psychological Science, 18*(4), 210-216.

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology, 65*, 71-93.

Kurtz, L. E., & Algoe, S. B. (2015). Putting laughter in context: Shared laughter as behavioral indicator of relationship well-being. *Personal Relationships, 22*(4), 573-590.

Langley, E. B. & Shiota, M. N. (in prep) Affective and social reactions to stereotype humor.

Latta, R. L. (1999). *The basic humor process*. Humor Research. Berlin and New York: Mouton de Gruyter.

Lerner, I., Bentin, S., & Shriki, O. (2012). Spreading activation in an attractor network with latching dynamics: Automatic semantic priming revisited. *Cognitive Science, 36*(8), 1339-1382.

Leveen, L. (1996). Only when I laugh: Textual dynamics of ethnic humor. *Melus, 21*(4), 29-55.

Lockyer, S., & Pickering, M. (Eds.). (2005). *Beyond a joke: The limits of humour* (p. 1). Basingstoke: Palgrave Macmillan.

Lu, A., Feng, Y., Yu, Z., Tian, H., Hong, X., & Zheng, D. (2015). Anxiety and mind wandering as independent consequences of stereotype threat. *Social Behavior and Personality: An International Journal, 43*(4), 537-558.

Maio, G. R., Olson, J. M., & Bush, J. E. (1997). Telling jokes that disparage social groups: Effects on the joke teller's stereotypes 1. *Journal of Applied Social Psychology, 27*(22), 1986-2000.

Martin, D. M. (2004). Humor in middle management: Women negotiating the paradoxes of organizational life. *Journal of Applied Communication Research, 32*(2), 147-170.

Martin, R. A., & Ford, T. (2018). The psychology of humor: An integrative approach. Academic press.

Martin, R. A., & Kuiper, N. A. (1999). *Daily occurrence of laughter: Relationships with age, gender, and Type A personality*.

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal Of Research in Personality, 37*(1), 48-75.

Martineau, W. H. (1972). A model of the social functions of humor. *The psychology of humor*, 101-125.

McBeath, M. K. (2018). Natural regularities and coupled predictive perceptual and cognitive biases: Why we evolved to systematically experience spatial illusions. In *Spatial biases in perception and cognition* (pp. 276-294). Cambridge University Press.

McGraw, A. P., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological Science, 21*(8), 1141-1149.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition.* Psychology Press.

Miller, S. S., O'Dea, C. J., Lawless, T. J., & Saucier, D. A. (2019). Savage or satire: Individual differences in perceptions of disparaging and subversive racial humor. *Personality and Individual Differences, 142*, 28-41.

Moè, A., Cadinu, M., & Maass, A. (2015). Women drive better if not stereotyped. Accident Analysis & Prevention, 85, 199-206.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal Of Personality and Social Psychology, 81*(1), 33.

Morrison, T. G., Morrison, M. A., McDonagh, L., Regan, D., & McHugh, S. J. (2014). Confirmatory factor and invariance analyses of the Motivation to Control Prejudiced Reactions Scale. *Open Journal of Statistics, 4*(6), 446-455.

Moskowitz, G. B. (2005). *Social cognition: Understanding self and others*. Guilford Press.

Necker, L.A. (1832). "Observations on some remarkable optical phaenomena seen in Switzerland; and on an optical phaenomenon which occurs on viewing a figure of a crystal or geometrical solid". *London and Edinburgh Philosophical Magazine and Journal of Science. 1* (5): 329–337. doi:10.1080/14786443208647909.

Neuberg, S. L., Williams, K. E., Sng, O., Pick, C. M., Neel, R., Krems, J. A., & Pirlott, A. G. (2020). Toward capturing the functional and nuanced nature of social stereotypes: An affordance management approach. In *Advances in experimental social psychology* (Vol. 62, pp. 245-304). Academic Press.

O'connor, E. C., Ford, T. E., & Banos, N. C. (2017). Restoring threatened masculinity: The appeal of sexist and anti-gay humor. *Sex Roles, 77*, 567-580.

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences, 118*(25), e2022340118.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of ClinicalEepidemiology, 49*(12), 1373-1379.

Piaget, J. (1970). *Genetic epistemology*. Columbia University Press.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107*(4), 786.

Pratt, S. B. (1998). Ritualized uses of humor as a form of identification among American Indians. InD. V. Tanno & A. Gonzalez (Eds.), *Communication and identity across cultures* (pp. 56–79). New York: Sage.

Provine, R. (2000). *Laughter:a scientific investigation*. NewYork: Viking.

R Core Team (2023). _R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raizada, R. (2020). A payload-ignition theory of adult-oriented humour: TUUTU (a Thought that is Unmentionable and Unmentioned, Triggered Unexpectedly).

Rappoport, L. (2005). *Punchlines: The case for racial, ethnic, and gender humor.* Greenwood Publishing Group.

Raskin, V. (1985). *Semantic Mechanisms of Humor*. Dordrecht, Boston, Lancaster: D. Reidel.

Raskin, V., Hempelmann, C. F., & Taylor, J. M. (2009). How to understand and assess a theory: The evolution of the SSTH into the GTVH and now into the OSTH.

Richeson, J. A., & Shelton, J. N. (2007). Negotiating interracial interactions: Costs, consequences, and possibilities. *Current Directions in Psychological Science, 16*(6), 316-320.

Riquelme, A. R., Carretero-Dios, H., Megías, J. L., & Romero-Sánchez, M. (2021). Joking for gender equality: Subversive humor against sexism motivates collective action in men and women with weaker feminist identity. *Sex Roles, 84*(1-2), 1-13.

Robinson, D. T., & Smith-Lovin, L. (2001). Getting a laugh: Gender, status, and humor in task discussions. *Social Forces, 80*(1), 123-158.

Romero-Sánchez, M., Durán, M., Carretero-Dios, H., Megías, J. L., & Moya, M. (2010). Exposure to sexist humor and rape proclivity: The moderator effect of aversiveness ratings. *Journal of Interpersonal Violence, 25*(12), 2339-2350.

Sailes, G. A. (1993). An investigation of campus stereotypes: The myth of Black athletic superiority and the dumb jock stereotype. *Sociology of Sport Journal, 10*(1), 88-97.

Saucier, D. A., Strain, M. L., Miller, S. S., O'Dea, C. J., & Till, D. F. (2018). "What do you call a Black guy who flies a plane?": The effects and understanding of disparagement and confrontational racial humor. *Humor, 31*(1), 105-128.

Saucier, D. A., O'Dea, C. J., & Strain, M. L. (2016). The bad, the good, the misunderstood: The social effects of racial humor. *Translational Issues in Psychological Science, 2*(1), 75.

Schneider, D. J. (2005). The psychology of stereotyping. Guilford Press.

Schoem, D. (2003). Intergroup dialogue for a just and diverse democracy. *Sociological Inquiry, 73*(2), 212-227.

Shapiro, J. P., Baumeister, R. F., & Kessler, J. W. (1991). A three-component model of children's teasing: Aggression, humor, and ambiguity. *Journal of Social and Clinical Psychology, 10*(4), 459-472.

Snefjella, B., Schmidtke, D., & Kuperman, V. (2018). National character stereotypes mirror language use: A study of Canadian and American tweets. *PloS One, 13*(11), e0206188.

Stangor, C., & Lange, J. E. (1994). *Mental representations of social groups: Advances in understanding stereotypes and stereotyping.* Academic Press.

Suls, J. M. (1972). A two-stage model for the appreciation of jokes and cartoons: an information processing analysis. In J. Goldstein & P. McGhee (Eds.), *The psychology of humor: theoretical perspectives and empirical issues* (pp. 81-100). New York: Academic Press.

Sutton, R. M., Elder, T. J., & Douglas, K. M. (2006). Reactions to internal and external criticism of outgroups: Social convention in the intergroup sensitivity effect. *Personality and Social Psychology Bulletin, 32*(5), 563-575.

Sydell, E. J., & Nelson, E. S. (2000). Modern racism on campus: A survey of attitudes and perceptions. *The Social Science Journal, 37*(4), 627-635.

Thai, M., Borgella, A. M., & Sanchez, M. S. (2019). It's only funny if we say it: Disparagement humor is better received if it originates from a member of the group being disparaged. *Journal of Experimental Social Psychology, 85*, 103838.

Veatch, T. C. (1998). *A theory of humor*.

Van Swol, L. M., & Sniezek, J. A. (2002). Trust me, I'm an expert: trust and confidence and acceptance of expert advice. In *8th International Conference on Behavioral Decision Research in Management (BDRAM), Chicago, Illnois, USA*.

Warren, C., Barsky, A., & McGraw, A. P. (2021). What makes things funny? An integrative review of the antecedents of laughter and amusement. *Personality and Social Psychology Review, 25*(1), 41-65.

Weaver, S. (2010). The 'Other'laughs back: Humour and resistance in anti-racist comedy. *Sociology, 44*(1), 31-48.

Weaver, R., Ferguson, C., Wilbourn, M., & Salamonson, Y. (2014). Men in nursing on television: exposing and reinforcing stereotypes. *Journal of Advanced Nursing, 70*(4), 833-842.

Xing, C. (2019). Group Membership, Content Valence, and Stereotype Agreement: Testing the Effects of Jokes and Asian Stereotypes (Doctoral dissertation, University of Kansas).

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes, 83*(2), 260-281.

APPENDIX A

UNIVERSITY APPROVAL FOR HUMAN SUBJECT RESEARCH

APPENDIX A

EXEMPTION GRANTED

On 10/6/2023 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Evaluations of Stereotype Humor
Investigator:	Michelle Shiota
IRB ID:	STUDY00018766
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• Consent, Category: Consent Form; • Debrief, Category: Other; • Images/Stimuli, Category: Other; • IRB protocol, Category: IRB Protocol; • Measures, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Recruitment, Category: Recruitment Materials;

The IRB determined that the protocol is considered exempt pursuant to Federal Regulations 45CFR46 (2)(i) Tests, surveys, interviews, or observation (non-identifiable) on 10/6/2023.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

If any changes are made to the study, the IRB must be notified at research.integrity@asu.edu to determine if additional reviews/approvals are required.