Understanding the Impact of Identity Threats on Hate Feelings: Exploring Identity Domains and Threat Types

Merve Mira Karakaya

S5276500

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Supervisor: Christian Martínez

Second evaluator: Mortada Al Amine

In collaboration with: Lieke Weerink, Jakob Rosenberg, Malin Schucher, Daniela Lofajova, Right de Vries.
June 20th, 2025

Abstract

In the digital era, identity is actively constructed and negotiated through online interactions, which can also expose individuals to identity threats. This study investigated how different types of identity threats (categorization and distinctiveness) and identity domains (demographics, personal preferences, and social values) affect the intensity of hate feelings toward an individual delivering the threat. Using a 3 (identity domain) x 2 (threat type) factorial design, 497 participants created personalized mock profiles and received tailored threatening messages simulating identity challenges in an online context. Contrary to the hypotheses, results showed no significant main or interaction effects of identity domain or threat type on hate feelings, measured with the Passionate Hate Scale. Qualitative responses suggested that one-off, impersonal threats from strangers may lack the psychological salience to evoke strong hate. Findings highlight the context-dependent nature of identity threat responses, particularly in low-stakes, interpersonal online interactions. The study contributes to understanding the complexities of emotional reactions to identity challenges and underscores the need for more immersive and realistic experimental approaches to explore identity threat and hate dynamics in digital environments.

Keywords: identity threat, hate feelings, identity domains, online interactions, interpersonal threat

A thesis is an aptitude test for students. The approval of the thesis is proof that the student has sufficient research and reporting skills to graduate, but does not guarantee the quality of the research and the results of the research as such, and the thesis is therefore not necessarily suitable to be used as an academic source to refer to. If you would like to know more about the research discussed in this thesis and any publications based on it, to which you could refer, please contact the supervisor mentioned.

Understanding the Impact of Identity Threats on Hate Feelings: Exploring Identity Domains and Threat Types

In the digital era, social media has become a central space for identity construction and self-presentation (Chen, 2023). Whether on platforms like Instagram and Twitter or dating apps such as Tinder, individuals carefully curate their online presence, selecting which aspects of themselves to highlight, through the content they share, the communities they engage with, or the personal details they disclose. This digital self-representation is more than just communication; it serves as a means of signaling belonging, distinguishing oneself from others, and reinforcing a personal or collective sense of identity (Chen, 2023; Ganda, 2024). On dating apps, for instance, users are often required to condense their identity into a few pictures and short bios, strategically choosing what aspects of themselves to emphasize in order to attract like-minded individuals (Degen & Kleeberg-Niepage, 2021; Yurchisin et al., 2005). These environments not only highlight the flexible and performative nature of identity expression, but also create spaces where identity-based confrontations can arise, making them a relevant and realistic setting for experimentally examining how people respond when their sense of identity is challenged.

However, the way people define themselves, both online and offline, is not random; it is shaped by multiple, intersecting factors that contribute to identity formation. According to Jones and McEwen's (2000) model, identity is a complex and multidimensional construct influenced by both social categorizations and personal experiences. Building on this framework, identity can be usefully explored through the lenses of demographic characteristics (such as race or gender), personal preferences (such as interests and tastes), and deeply held social values (such as political or moral beliefs). These dimensions help clarify how individuals understand themselves and relate to others in various social contexts. Demographic identity includes characteristics such as gender, nationality, and ethnicity, which are often treated as stable and externally assigned, while frequently influencing how individuals are categorized by society (Clair et al., 2018; Josselson & Harway, 2012; Saperstein, 2024). Preferences reflect personal tastes and interests that tend to be flexible and self-determined. Research shows that these preferences are actively explored during key developmental stages like emerging adulthood and contribute to identity formation (Schwartz et al., 2013). For example, preferences may include interests in certain activities, lifestyle choices, or cultural tastes. Social values, by contrast, encompass deeply held ideological beliefs, including political and moral worldviews, which guide how individuals align with broader cultural and societal movements. Values have been identified as a core component of personal identity, shaping motivations, behaviors, and social affiliations (Hitlin, 2003). Among other identity sources, these three domains collectively seem to form a good representation of individuals' identity, influencing not only how individuals see themselves but also how they engage with and are perceived by others.

While identity provides a sense of stability and belonging, individuals often encounter situations where their identity is challenged, misrepresented, or constrained by external forces (Branscombe et al., 1999). Research suggests that when people perceive their social identity to be under threat, it can lead to heightened intergroup tensions and hostility. Across different cultural contexts, symbolic threats, perceived challenges to one's values or way of life, have been shown to strongly predict negative attitudes toward outgroups (Obaidi et al., 2018). Moreover, Martínez et al. (2022) found that symbolic threats significantly predicted hate in both

interpersonal and intergroup settings. In their threat-based hate model, they showed that such threats not only increase hate but also lead to greater support for aggressive behaviors. They also found that unlike other negative emotions such as anger or dislike, hate is more emotionally arousing, personally threatening, and strongly linked to attack-oriented tendencies. This view aligns with Sternberg's (2008) Triangular Theory of Hate, which conceptualizes hate as a complex emotion involving negation of intimacy, passion, and commitment, reflecting emotional intensity, relational rejection, and cognitive reinforcement. This suggests that when people perceive threats to who they are, they may experience hate as a distinct and powerful emotional response. Similarly, feelings of collective disadvantage and identity threats have been linked to stronger group-based responses, such as increased ethnocentrism and a greater willingness to engage in collective action (Grant & Brown, 1995). These patterns suggest that identity threats are not merely personal but can have serious social consequences, shaping how groups and people relate to one another. However, they do not always take the same form. Some threats challenge deeply held beliefs, while others can undermine a person's sense of distinctiveness or belonging. One influential framework for understanding these dynamics comes from Branscombe et al. (1999), who identifies four types of identity threats, whereas this study will be focusing on two in particular: categorization threats and distinctiveness threats. Categorization threats arise when individuals are placed into broad or unwanted social categories, stripping them of individuality and imposing characteristics they may not identify with. These threats are often amplified in online spaces, where users are frequently stereotyped based on limited cues such as gender, nationality, or political beliefs. In fragmented and high-choice media environments, as Arendt (2023) notes, people are increasingly exposed to content that aligns with their preexisting views. This self-selected exposure can reinforce stereotype-consistent narratives and perceptions, making online interactions more susceptible to simplistic, prejudiced categorizations. As a result, users may not only be mischaracterized by others, but also internalize or anticipate reductive judgments, heightening the impact of identity-based threats (Arendt, 2023; Onyango & Bowe, 2019). In contrast, distinctiveness threats occur when the uniqueness of a group to which an individual belongs to is undermined, making it harder to maintain a clear and differentiated social identity.

It is important to look at these dynamics as they reflect how identity confrontations impact individuals' psychological responses and social behaviors. Although prior research has explored identity threats, much of the existing literature focuses on symbolic threats, which broadly encompass threats to identity, values, and culture (Martínez et al., 2022; Makashvili et al., 2018). However, less attention has been given to how specific identity domains, demographics, preferences, and social values, differ in their psychological impact when threatened. This distinction is crucial because identity is not a monolithic concept; different aspects of identity may provoke distinct emotional and behavioral responses when challenged.

Research suggests that self-chosen identity aspects, such as preferences and social values, may be more central to individuals' sense of self, and as a result, people could be more sensitive to threats in these areas. However, although Easterbrook et al. (2020) found that social class–based identities (e.g., education, occupation) are often more central than immutable demographic traits, there is still limited research directly comparing the centrality of preferences and values to demographics. Therefore, the study will maintain an explanatory approach, exploring these potential relationships rather than asserting them definitively. Moreover, while identity threats have been examined in relation to prejudice and intergroup conflict, little research has focused on their direct impact on individuals, particularly regarding the hate they may feel as a result. In addition, whereas most existing work focuses on intergroup-level dynamics, this study adopts an interpersonal lens to examine how individuals react emotionally when their identity is challenged not by a group, but by a single voice. Even though the threat comes from an individual, it still targets socially significant aspects of identity, allowing us to explore how social identity processes manifest in one-on-one interactions. This shift in focus reflects the reality that identity threats do not always come from collectives, sometimes, it takes only one person to unsettle the self. Even in brief, impersonal exchanges, the self can remain sensitive to cues of social meaning, especially when they touch on personally meaningful dimensions of identity.

Lastly, there is a lack of comparative research on the different types of identity threats, such as categorization and distinctiveness threats, and their respective impacts. Understanding these dynamics is important, especially in digital environments, where identity confrontations are frequent. This gap in the literature presents an opportunity to further explore how these threats contribute to hate, and their broader implications for online hostility, interpersonal conflict arising from identity threats, and polarization.

Based on the theoretical framework of identity threats and emotional responses (Branscombe et al., 1999), this study examines how different types of identity threats influence the intensity of hate feelings when targeting one's identity. Specifically, it tests the following hypotheses:

Hypothesis 1. The intensity of hate feelings toward the individual delivering the threat will vary depending on the identity domain that is targeted (demographics, preferences, or social values).

Hypothesis 2. The intensity of hate feelings toward the threatener will vary depending on the type of identity threat experienced (categorization vs. distinctiveness).

Hypothesis 3. The intensity of hate feelings will depend on the interaction between the identity domain targeted and the type of threat (categorization vs. distinctiveness) delivered by the individual.

Methods

Participants

An a priori power analysis was conducted to determine the required sample size for a model suitable for testing the hypotheses within a larger project, including a 3 (identity type) x 3 (threat type), and one covariate. The analysis intended to detect medium effect sizes ($f = 0.25$), with an alpha level of .05 and a desired power of 0.9. Based on this, we planned to recruit $N = 330$ participants. Due to technical setbacks with the randomization, we increased the sample size to $N = 608$ to ensure an equal number of participants in each condition, and dropped the control no threat condition due to imbalance in the cells. Four participants were excluded after withdrawing their consent, 14 failed the attention check and four participants were removed because they gave identical descriptions in the personal prompt. Of the final sample ($N = 497$) recruited through Prolific, 250 identified as female (50.3%), 244 as male (49.1%) and 3 as other (0.6%). The participants ranged in age from 18-81, with a mean age of 39.25 years ($SD = 13.56$). All participants were citizens of the United States of America, and the sample was pre-screened to be balanced by sex and partisanship (M political orientation = 0.170, on a scale from -1 to 1).

Instruments

Hate

The assessment of Hate was conducted with an adaptation of the Passionate Hate Scale developed by Zeki and Romaya (2008). Drawing on the triangular theory of hate (Sternberg, 2008), the scale measures three core components: anger, disgust, and contempt. Minor adjustments were made to check attention, and three items were reverse coded, one for each subscale. In total, the scale involved 13 items, including 4 items for each subscale and one attention check question. Examples of the items are “I cannot control my anger towards this kind of person” (anger), “this kind of person is really distrusting” (disgust) and “the world would be a better place without this kind of person” (contempt). These items were rated on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

Manipulation Materials

Before the experimental manipulation, participants created a mock identity profile designed to simulate a social media–style introduction. They first selected a fictional nickname and responded to a bio prompt describing how someone close to them would characterize them. Then, they selected and ranked identity-relevant “badges” across three domains: personal characteristics (e.g., age, sex, ethnicity), hobbies and lifestyle preferences (e.g., music, fitness, creativity), and values and allyships (e.g., political ideology, social causes, dietary values). For each top-ranked badge, participants answered follow-up questions specifying their identity (e.g., “What is your political ideology or affiliation?”). At the end of this phase, participants reviewed a summary of their profile before confirming it.

Participants were told that their profile would be reviewed by another player in an online matching game and that they would receive feedback from that person. In reality, all participants were randomly assigned to one of six experimental conditions in a 3 (identity domain: demographics, preferences, values) × 2 (threat type: categorical, distinctiveness) between-subjects design. In each condition, participants received a negative, identity-threatening message, allegedly written by another participant, that targeted the identity domain they had emphasized in their profile. These messages were designed to devalue a personally meaningful aspect of the participant’s self-concept, such as their values, preferences, or demographic traits. For instance, participants in the values–distinctiveness threat condition received a message implying that their values were not unique or meaningful. A control group received a neutral, positive message instead. This manipulation was designed to simulate an identity-relevant threat in a controlled yet ecologically valid way.

Procedure

Data were collected through an online Qualtrics survey using a convenience sample of paid U.S.-based participants. After providing informed consent, participants entered a self-chosen nickname and completed a mock identity profile. In this profile, they answered questions about aspects of their identity they considered central (e.g., gender, politics, ethnicity, sports), spanning three domains: demographics, preferences, and values. Participants were then told they would be matched with another person for a brief interaction. In reality, all responses were pre-

generated. Based on the identity domain they had emphasized most, participants in the experimental condition received an identity-threatening message, allegedly written by another participant that threatened their identity through two mechanisms: by categorizing them into a generalized group (categorization threat), or by minimizing the uniqueness of their identity (distinctiveness threat). Those in the control group received a neutral message. The study was approved by the Ethics Committee of the University of Groningen, and the survey took approximately 18 minutes to complete on average.

Design and Analytical Strategy

This study used a randomized experimental, cross-sectional design with a 3 (identity domain: demographics, preferences, values) × 2 (threat type: categorization vs. distinctiveness) between-subjects factorial structure, resulting in six experimental conditions. Participants were randomly assigned to one of the six conditions. The manipulation was tailored to target the identity domain participants had emphasized in a prior mock profile task. A control group receiving a neutral message was included but later dropped due to cell imbalance.

To test the hypotheses, a two-way analysis of variance (ANOVA) was conducted. The independent variables were identity domain and threat type. The dependent variable was the intensity of hate feelings, measured using the adapted Passionate Hate Scale (Sternberg, 2008; Zeki & Romaya, 2008). This analysis allowed for testing both main effects of identity domain and threat type, as well as the interaction between them.

Prior to analysis, assumption checks were conducted, including the Shapiro–Wilk test for normality and Levene's test for homogeneity of variances. In the case of a significant interaction effect, post-hoc analyses using Bonferroni-adjusted pairwise comparisons were planned to explore differences between specific condition pairs.

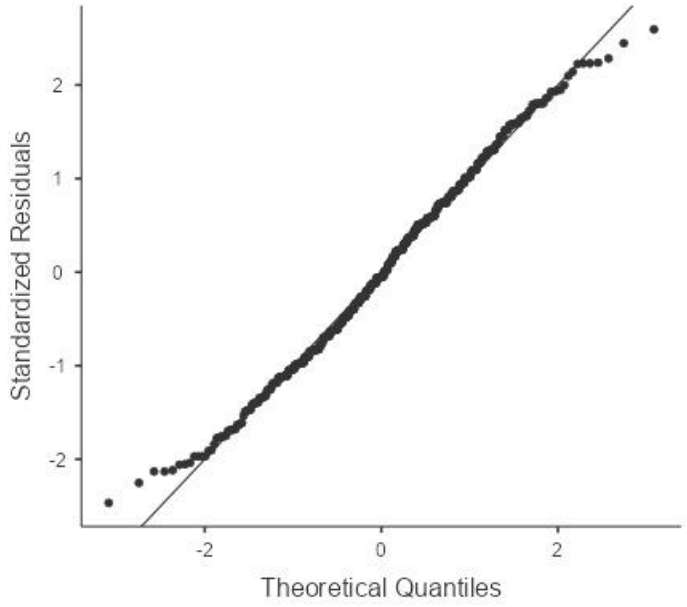
Results

A 3 (identity domain: demographics, preferences, social values) × 2 (threat type: categorization vs. distinctiveness) factorial ANOVA was conducted to test the study's hypotheses. Assumption checks indicated that the homogeneity of variances was met (Levene's test, $p > .05$). Although the Shapiro-Wilk test suggested a violation of the normality assumption ($p < .05$), visual inspection of the Q–Q plot showed that standardized residuals closely followed a straight line, indicating only minimal deviation (see Figure 1). Given this, and the robustness of ANOVA to such violations in large samples, the analysis proceeded as planned. The ANOVA revealed no significant main effects or interaction effects, all p-values > .05 (see Table 1). Effect sizes were small (partial $\eta^2 s < .01$), suggesting minimal differences across experimental conditions.

Although the effects were not statistically significant, small descriptive differences in hate scores were observed across the six conditions. The preferences + distinctiveness and preferences + categorization conditions showed the highest average hate scores (*M* = 3.99, *SD* = 1.09 and *SD* = 1.24, respectively), followed by demographics + distinctiveness and demographics + categorization (both *M* = 3.97, *SD* = 1.24 and *SD* = 1.19). Lower scores were found in the values + categorization condition (*M* = 3.81, *SD* = 1.12) and the values + distinctiveness condition (*M* = 3.80, SD = 1.18) (see Figure 2A, Appendix A). While these differences are small, they may hint at subtle differences in how identity threats are experienced across domains, but should be interpreted with caution due to the lack of significant effects.

To complement the quantitative findings, participants' open-ended responses were informally categorized into four groups based on tone and content. Out of approximately 160 replies, 70 were helpful or positive, reflecting respectful, supportive, or explanatory comments (e.g., *"I hope your day gets better buddy," "Get to know people before you judge them," "It is good to be considerate and accept that different people will have different opinions"*). About 25 responses were somewhat helpful or neutral, including polite but brief or minimally engaging replies (e.g., *"Good one," "Thank you for your opinion," "Your opinion is appreciated"*). Notably, 55 responses were negative in tone, but these varied in severity. Of these, 26 were explicitly hateful, containing direct insults, aggressive language, or hostile remarks (e.g., *"You need therapy," "I'll beat your ass boy," "You're a disappointment to humanity"*). The remaining 29 negative replies were unhelpful or dismissive, characterized by sarcasm, rudeness, or dismissiveness without outright hateful content (e.g., *"Whatever, your opinion doesn't matter," "Cool story, bro," "Not surprised by your ignorance"*). A small number of responses (around 5) were unclear or irrelevant (e.g., *"A great deal," "Hmm"*). This nuanced categorization provides richer context to understand the emotional and cognitive reactions participants had to the identity-threatening messages, highlighting a spectrum of engagement from constructive dialogue to hostility.

Figure 1
Q–Q Plot of Standardized Residuals Showing Approximate Normality



Note. The points follow the diagonal line closely, suggesting that the residuals meet the assumption of normality.

Table 1

Results of 3 × 2 Factorial ANOVA on Hate Scores

ANOVA - hate total				
	Sum of Squares	df	Mean Square	F
p				
Id_condition	3.49533	2	1.74766	1.25984
0.285				
Threat_condition	0.00529	1	0.00529	0.00381
0.951				
id_condition * Threat_condition	0.00594	2	0.00297	0.00214
0.998				
Residuals	681.11769	491	1.38721	

Note. Results of the 3 (identity domain: demographics, preferences, social values) × 2 (threat type: categorization vs. distinctiveness) factorial ANOVA on hate scores. No significant main or interaction effects were found (all $p > .05$). Effect sizes were small (partial $\eta^2 < .01$), indicating minimal differences between conditions.

Discussion

Summary of Findings

This study set out to explore how identity threats, specifically categorization and distinctiveness threats, impact hate feelings when different identity domains (demographics, preferences, or social values) are targeted. Drawing from theories of identity threat (Branscombe et al., 1999) and the triangular theory of hate (Sternberg, 2008), we hypothesized that both the type of identity domain and threat would influence the intensity of hate experienced by participants toward an individual who challenged their identity. However, contrary to our expectations, the results revealed no statistically significant main effects or interaction effects. In other words, the identity domain targeted and the type of threat posed did not significantly influence the level of hate participants felt.

Interpretation of Results

These findings raise several questions. Why, despite strong theoretical backing, did we fail to detect significant differences? One explanation may lie in the interpersonal nature of the threat. Much of the literature linking identity threats to hate has examined intergroup contexts (e.g., political or cultural conflicts), where the threat is perceived as coming from an opposing group (Obaidi et al., 2018; Martínez et al., 2022). In contrast, participants in our study received a single critical message from a supposed peer, essentially a stranger in an online interaction. Without an ongoing relationship or strong group affiliation to contextualize the interaction, the threat may have lacked psychological salience. That is, participants may not have felt especially violated by someone they did not know or care about. Unlike intergroup hate, which is often fueled by collective narratives and perceived historical injustice (Bar-Tal, 2011), interpersonal hate may require a baseline level of relational investment or perceived betrayal, neither of which were present in our design.

Additionally, our sample did not appear particularly reactive to the threat. Descriptive statistics on the Passionate Hate Scale showed relatively low to moderate scores overall (Ms around 3.8–4 on a 7-point scale), suggesting that participants did not feel especially hateful. This may indicate that the manipulation was not strong enough to induce a meaningful emotional response. Although the messages were designed to undermine meaningful aspects of identity, they may have come across as generic or impersonal. Our informal content analysis of open-ended responses supports this interpretation: rather than expressing hate, many participants responded with attempts to explain themselves, offered appeals, or used humor or sarcasm to deflect the threat. This suggests that the manipulations may not have successfully conveyed a credible or emotionally impactful threat.

Another explanation concerns the centrality of identity domains. While Easterbrook et al. (2020) suggest that social class–based identities may be more central to individuals' sense of self than immutable demographics, the specific centrality of values and preferences relative to demographics remains less clear in the literature. Interestingly, our findings showed the highest average hate scores in preference-related threat conditions, whereas value-related threats yielded slightly lower scores. This contrasts with the expectation that social values would provoke stronger emotional reactions when threatened. One possibility is that participants felt greater ownership over their preferences (e.g., hobbies, interests), as these are more individualized and self-chosen, while values might be seen as more abstract or socially shared. Alternatively, people may be more accustomed to having their values challenged in public discourse, especially in polarized environments, and thus feel less personally attacked when this occurs.

Similarly, the lack of significant differences between categorization and distinctiveness threats may point to the subtlety of the distinction as it played out in this design. Although theoretically distinct, these types of threats may not have been easily distinguishable to participants based on a single sentence of feedback. Moreover, the brief, text-based nature of the manipulation, delivered in an artificial, game-like format, may not have adequately captured the emotional nuance or gravity typically associated with identity threat. Future studies might benefit from using more immersive or interactive scenarios, perhaps through role-playing, video messages, or repeated exchanges to better simulate real-world identity confrontations.

Theoretical and Practical Implications

Although this study found no significant effects of identity threat types or domains on hate feelings, the results offer valuable insight into the complexity of emotional responses to identity threats in online interpersonal contexts. Identity threats do not always trigger strong negative emotions, especially when interactions are brief, anonymous, and involve strangers without ongoing relationships. This underscores the importance of personal relevance and relational investment in shaping whether threats feel meaningful or hurtful. Most identity threat theories focus on group-level or prolonged conflicts with higher emotional stakes (Branscombe et al., 1999; Obaidi et al., 2018), but our findings suggest that in low-stakes, one-off online encounters, emotional impact may be muted because individuals lack personal connection or accountability. Thus, emotional intensity depends not only on the threat content but also on context, such as threat credibility and relational dynamics. These nuances highlight the need for models of identity threat to more explicitly incorporate situational and relational factors.

Additionally, the lack of significant differences in hate feelings between categorization and distinctiveness threats suggests these may not be psychologically distinct in low-stakes online settings. While categorization threats involve broad group labels and distinctiveness threats involve loss of uniqueness within a group (Branscombe et al., 1999), participants did not differentiate their emotional responses. Without strong relationships or ongoing social context, the subtle distinctions between these threats may blur where the clear "signal" of threat type is drowned out by the "noise" of anonymity and detachment. In such minimal connection and consequence, people may simply register a general sense of "threat" rather than nuanced

emotional reactions.

From a practical standpoint, the finding that brief, anonymous identity threats in online settings generally do not provoke strong hate feelings suggests that platform moderators and designers might prioritize repeated or personally targeted threats, which are more likely to cause sustained emotional harm. Isolated, low-stakes incidents often elicit milder reactions and may not require intensive intervention. However, this does not imply that all brief threats are harmless. A notable portion of participant responses to identity-threatening prompts contained explicit hostility, demonstrating that even brief, anonymous interactions can sometimes trigger emotionally charged exchanges. This underscores the importance of providing users with easily accessible tools, such as blocking or muting features, to manage negativity promptly. Enhancing the visibility and usability of these tools, especially near negative content, can empower users to de-escalate potential conflicts before emotional responses escalate.

Strengths and Limitations

In terms of strengths, the study had several notable features. First, the large sample size ($N = 497$) provided sufficient statistical power to detect small to medium effects, increasing the reliability of the results. Second, the study employed a novel manipulation method, a mock social media-style identity profile, which allowed participants to self-select identity elements that felt meaningful to them, adding ecological validity. Third, the use of the Passionate Hate Scale (Zeki & Romaya, 2008) offered a multidimensional measure of hate, capturing not just anger but also contempt and disgust, making the emotional response more nuanced.

Nonetheless, the study also faced several limitations. The manipulation may have been too weak to elicit strong emotional reactions, as discussed. Moreover, the study lacked a meaningful control group due to cell imbalance, which limited our ability to compare the experimental conditions against a baseline response. Finally, the experimental setting may not have felt realistic or intimate enough for participants to feel personally attacked, especially given the anonymity and impersonality of online communication. This may have reduced the emotional intensity of the manipulation and introduced reactivity bias, where participants behave more “rationally” or cautiously in response to an artificial situation.

Future Directions

Given these findings, there are several future directions for research. First, future studies should aim to increase the salience and realism of identity threats, perhaps by incorporating multimedia stimuli or real-time interaction with confederates. Second, researchers could explore longitudinal or repeated-interaction designs, where identity threats accumulate over time, better simulating real-world online environments. Third, further research could investigate the boundary conditions of interpersonal hate: what makes a one-time comment from a stranger provoke hate in some contexts but not others? Exploring individual difference factors, such as identity centrality, sensitivity to rejection, or past experiences of discrimination, could provide richer insights.

In conclusion, while this study did not confirm our original hypotheses, it contributes to the growing literature on identity threat by highlighting the complex and context-dependent nature of emotional responses to interpersonal identity challenges. As a pilot exploration, it emphasizes the need for more nuanced, immersive approaches to understanding how identity threats unfold in digital interactions and how they may (or may not) give rise to hate.

Conclusion

This study set out to explore how different types of identity threats, categorization and distinctiveness, across various identity domains (demographics, preferences, and values) affect feelings of hate in interpersonal contexts. While our hypotheses were not supported by statistically significant results, the findings offer valuable insight into the emotional responses people display when their self-concept is challenged, even by a stranger. The absence of strong hate reactions, paired with participants’ reflective and sometimes even prosocial open-ended responses, suggests that the experience of interpersonal identity threat may not always provoke intense negative emotions, particularly in contexts lacking personal investment. Our findings point to several possibilities: that the intensity of the manipulation may have been insufficient to evoke strong hate; that participants may have compartmentalized the threat as unimportant due to the anonymous, non-intimate context; or that modern online dynamics encourage more emotionally filtered responses. It may be that in digital spaces, where interactions are brief and accountability is low, people develop a kind of emotional distance, a filter that allows them to engage without fully investing. When identity threats occur in these conditions, the self may still notice the challenge, but not feel compelled to respond with intensity. The emotional cost simply isn’t worth it. In that sense, muted reactions might reflect more than apathy; they could be a kind

of strategic indifference, an adaptive way of protecting the self in environments where social bonds are weak and stakes are low. This pilot study, while limited in scope, contributes to the growing understanding of how identity processes unfold in digital environments. It highlights the need to refine experimental manipulations and points toward the importance of context, relationship dynamics, and personal relevance in shaping responses to identity threats. Ultimately, the findings suggest that the self is not only shaped by what threatens it, but by what it chooses to respond to. Though the effects were small and non-significant, the study opens the door to further investigation into how identity confrontations are emotionally navigated in a world increasingly defined by online interactions.

References

Arendt, F. (2023). Media stereotypes, prejudice, and preference-based reinforcement: toward the dynamic of self-reinforcing effects by integrating audience selectivity. *Journal of Communication*, *73*(5), 463–475. https://doi.org/10.1093/joc/jqad019

Bar-Tal, D. (2011). *Intergroup Conflicts and Their Resolution*. Psychology Press.

Branscombe, N. R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. In R. Spears, P. J. Oakes, N. Ellemers, & S. A. Haslam (Eds.), Social identity: Context, commitment, content (pp. 35–58). Blackwell Publishers. https://archive.org/details/1999-branscombeetal.-soc-id-threat

Chae, M. H. (2001). Gender and ethnicity in identity formation. *New Jersey Journal of Professional Counseling*. https://www.researchgate.net/publication/339314450_Gender_and_Ethnicity_in_Identity_Formation-New_Jersey_Journal_of_Professional_Counseling

Chen, S. (2023). The influence of social media platforms on Self-Identity in the new media environment: the case of TikTok and Instagram. SHS Web of Conferences, *165*, 01020. https://doi.org/10.1051/shsconf/202316501020

Clair, J. A., Humberd, B. K., Rouse, E. D., & Jones, E. B. (2018). Loosening categorical thinking: extending the terrain of theory and research on demographic identities in organizations. *Academy of Management Review*, *44*(3), 592–617. https://doi.org/10.5465/amr.2017.0054

Degen, J. L., & Kleeberg-Niepage, A. (2021). Profiling the Self in Mobile Online Dating Apps: a Serial Picture Analysis. *Human Arenas*, *6*(1), 147–171. https://doi.org/10.1007/s42087-021-00195-1

Easterbrook, M. J., Kuppens, T., & Manstead, A. S. R. (2019). Socioeconomic status and the structure of the self-concept. British Journal of Social Psychology, 59(1), 66–86. https://doi.org/10.1111/bjso.12334

Ganda, M. (2014). Social Media and Self: Influences on the Formation of Identity and Understanding of Self through Social Networking Sites. https://doi.org/10.15760/honors.64

Grant, P. R., & Brown, R. (1995). From ethnocentrism to collective protest: responses to relative deprivation and threats to social identity. *Social Psychology Quarterly*, *58*(3), 195. https://doi.org/10.2307/2787042

Hitlin, S. (2003). Values as the Core of Personal Identity: Drawing Links between Two Theories of Self. *Social Psychology Quarterly*, *66*(2), 118. https://doi.org/10.2307/1519843

Jones, S. R., & McEwen, M. K. (2000). A conceptual model of multiple dimensions of identity. *Journal of College Student Development, 41*(4), 405–414. https://www.researchgate.net/publication/292759031_A_conceptual_model_of_multiple_dimensions_of_identity

Josselson, R., & Harway, M. (Eds.). (2012). *Navigating multiple identities: Race, gender, culture, nationality, and roles*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199732074.001.0001

Makashvili, A., Vardanashvili, I., & Javakhishvili, N. (2018). Testing intergroup threat theory: Realistic and symbolic threats, religiosity and gender as predictors of prejudice. *Europe's Journal of Psychology*, *14*(2), 464–484. https://doi.org/10.5964/ejop.v14i2.1483

Martínez, C. A., Van Prooijen, J., & Van Lange, P. a. M. (2022). Hate: Toward understanding its distinctive features across interpersonal and intergroup targets. *Emotion, 22*(1), 46–63. https://doi.org/10.1037/emo0001056

Obaidi, M., Kunst, J. R., Kteily, N., Thomsen, L., & Sidanius, J. (2018). Living under threat: Mutual threat perception drives anti-Muslim and anti-Western hostility in the age of terrorism. *European Journal of Social Psychology, 48*(5), 567–584. https://doi.org/10.1002/ejsp.2362

Onyango, J. M., & Bowe, S. N. (2019). Seeing is believing: How social media is challenging physician stereotypes. *Journal of Graduate Medical Education*, *11*(5), 495–497. https://doi.org/10.4300/jgme-d-19-00266.1

Saperstein, A. (2024). Recognizing identity fluidity in demographic research. *Population and Development Review*. https://doi.org/10.1111/padr.12670

Schwartz, S. J., Zamboanga, B. L., Luyckx, K., Meca, A., & Ritchie, R. A. (2013). Identity in emerging adulthood. *Emerging Adulthood*, *1*(2), 96–113. https://doi.org/10.1177/2167696813479781

Sternberg, R. J., & Sternberg, K. (2008). The duplex theory of hate I: The triangular theory of the structure of hate. In *The nature of hate* (pp. 51–77). Cambridge University Press. https://doi.org/10.1017/cbo9780511818707.004

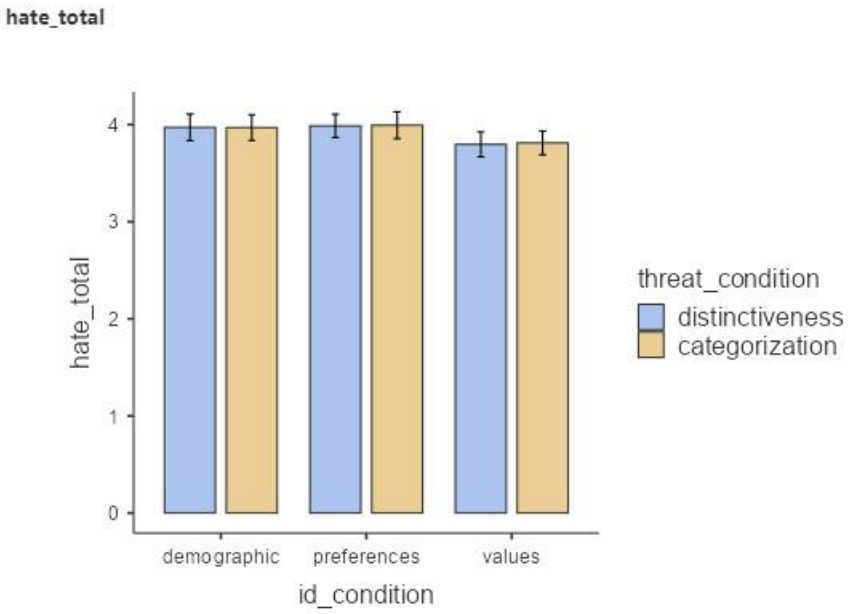
Yurchisin, J., Watchravesringkan, K., & McCabe, D. B. (2005). An exploration of identity re-creation in the context of internet dating. Social Behavior and Personality: *An International Journal, 33*(8), 735–750. https://doi.org/10.2224/sbp.2005.33.8.735

Zeki, S., & Romaya, J. P. (2008). Neural correlates of hate. *PLoS ONE, 3*(10), e3556. https://doi.org/10.1371/journal.pone.0003556

Appendix A

Figure 2A

Bar chart of mean scores across conditions



Note. The bar chart shows average hate scores for each combination of identity domain (demographics, preferences, social values) and threat type (categorization, distinctiveness). The preferences conditions had the highest mean scores (~4.0), followed closely by demographics conditions (~3.97), with values conditions showing the lowest means (~3.80). Error bars represent standard deviations. Although no significant differences were found in the ANOVA, these descriptive trends may suggest subtle variation in hate responses across conditions.

Appendix B

Declaration of AI Use

I have not presented any AI-generated content as my own original work. I used ChatGPT solely to help improve grammar, clarify wording, and refine my explanations while drafting this paper.

The following prompts were input into ChatGPT:

- "What changes can I make to this sentence to enhance its style and coherence?"
- "Explain this concept in simpler terms."
- "Does this paragraph flow well, or does it need restructuring?"

The outputs were reviewed and modified by me to fit the context and academic style of my paper.