Comment like Nobody's Watching: Perceived Anonymity in Online Spaces

Sacha Bernardus, S2841339

Faculty of Behavioural Management and Social Sciences, University of Twente

2024-202000308-2A: M12 Bachelor thesis COM

Roel Lutkenhaus

June 30, 2025

Table of Contents

Abstract

This study aimed to answer the question: 'To what extent does perceived anonymity have an effect on norm-breaking behaviour in an online comment section?'. Building off of previous research on Norm Theory, this study tried to understand the effects of perceived anonymity in online comment sections on the likelihood of someone showing norm-breaking behaviour and if the perceived norms within a comment section have an effect on that relationship. Using a between-subjects experimental design with 4 conditions, 83 participants were asked to leave comments on a fake website made to look like TikTok. Each participant was assigned to a different condition, which changed what type of comments they saw next to the videos. Comments could either be positive or negative, and they can be left by perceived anonymous users or very identifiable users. The comments left by participants were coded on their levels of Profanity, Hostility, Inappropriateness and Community Guideline Violations. The results show that negative commenting environments show higher levels of norm-breaking behaviour, regardless of the level of perceived anonymity. A post-experiment survey also shows that most people dislike profanity and violations, but are alright with sarcasm and insensitive behaviour. These results suggested that perceived social norms have a stronger effect on the likelihood of norm-breaking behaviour occurring than anonymity alone. This study offered insights into the social dynamics within online comment sections and highlighted the importance of moderation and tone setting in these environments.

Introduction

Imagine watching a video on YouTube about the Barbie movie or watching a TikTok about a new trend. The content itself can be harmless, but when you open the comment section, you stumble upon a world of uncontrolled chaos. The comments are tearing the creator apart, taking part in heated debates and insulting each other left and right. When you check to see who would say such outlandish things, you notice that most of the comments are placed by users with no profile picture and a username similar to 'User12345'. These comments are placed by users who seem to be acting without any kind of restraint due to their perceived anonymity. At first glance, it seems that anonymity doesn't just remove someone's name and photo but also removes any social norms they would normally adhere to.

Norm theory explains how people behave in certain contexts where certain social norms apply and how these norms are established. This study will mainly be building off of Norm Theory and Suler's (2004) online disinhibition effect. As opposed to laws, social norms are not always clearly written down or defined but instead are discovered, created and learned through social interaction or observation (Rimal & Lapinski, 2015). Different fields of research will define social norms slightly differently, so this paper will give an overview of the norms that are relevant for this study and how they are defined. According to a study done by Cialdini et al. (1991), there are two main types of social norms, descriptive norms and injunctive norms. Descriptive norms, as described by Cialdini et al. (1991), are the norms that are created by looking at what is most typically done by others. It is based on what you can see others do. Injunctive norms are based on what a person perceives as typically approved of or disapproved of by others. It is based on what a person thinks others would find okay and not okay to do. Suler's (2004) theory expands on this concept by adding some psychological factors, including perceived anonymity. Suler highlights how these factors can reduce the

ability to hold back inappropriate or unwanted behaviour and increase the likelihood of antisocial behaviour online. Together, these theories create a strong framework which can be used to investigate online norm-breaking behaviour.

This study aims to fill in a research gap in existing literature by investigating the relationship between perceived anonymity and norm-breaking behaviour within online comment sections on popular social media platforms. While previous research has laid the foundation in understanding online environments and how social norms shape our behaviour, these theories have not often been applied to the platform-specific and interactive context of comment sections. Researching the effect of perceived anonymity in this context could provide helpful insights into the psychological side of user behaviour and could help create more effective moderation tactics.

The ever-rising popularity of social media platforms has made it extremely easy for anyone to participate in online discussions. Online spaces can change what is considered 'normal' to say and what is not when compared to real-life situations, and despite the best efforts of moderation teams on these online platforms, norm-breaking behaviour is still extremely prevalent. These kinds of behaviours can have positive outcomes by creating a safe space for people to anonymously share their struggles and feelings, but they can also foster an environment filled with hostility and harmful behaviours. That is why this study addresses the following research question: 'To what extent does perceived anonymity have an effect on norm-breaking behaviour in an online comment section?'. The objective of this study is to estimate how perceived anonymity in social media users affects their likelihood of violating social norms in the context of a comment section.

Theoretical Framework

To start answering the question about the effect of perceived anonymity on norm-breaking behaviour in online comment sections, some core concepts need to be defined: anonymity, norm-breaking behaviour and social norms in online spaces. By comparing previous studies and pulling them into a modern context, it is possible to gain a deeper understanding of the variables in play and what they mean in the context of this specific study.

Defining Anonymity

Anonymity in online contexts can be hard to define since complete online anonymity is almost unachievable in modern times. When a user does not have a picture or name on their profile, it is still possible for a tech-savvy person to track down someone's IP address. The IP address can show exactly what device is placing these comments and where that device is located. Marx (1999) states, "To be fully anonymous means that a person cannot be identified according to any of the seven dimensions of identity knowledge ." (p. 100). The seven dimensions that Marx presents are legal name, locatability, pseudonyms linked to name or location, pseudonyms that are not linked to name or location, pattern knowledge, social categorisation and symbols of eligibility/ineligibility. When a person manages to hide their name, location, patterns of behaviour and any other identifying traits that they might have, they can achieve anonymity.

This study on online spaces, however, focuses only on perceived anonymity, which refers to the feeling of being anonymous, no matter how anonymous a person truly is (Joinson, 1999). Unlike technical anonymity, in the form of IP addresses or similar account information, perceived anonymity focuses on how much a person believes they are anonymous. This same study also highlights how perceived anonymity influences online behaviour and how users act differently when they believe they cannot be identified. What is

clear from existing literature is that anonymity is never dichotomous; there are multiple types of anonymity. A study by Scott (2004) clearly distinguishes two forms of anonymity: physical and discursive. Physical anonymity refers to the physical source of a message being unknown, for example, by hiding your identity through appearance, like with a disguise. It focuses on hiding one's identity in the real, physical world. Discursive anonymity refers to the ability to remain anonymous in the realm of communication (like online spaces) by preventing others from identifying you based on what you say. By taking on a pseudonym in the form of a username and using a profile photo that does not include your face, one can achieve discursive anonymity.

Defining Norm-Breaking Behaviour

Norms are shared expectations and standards regarding appropriate behaviour in certain social contexts. Kahneman and Miller (1986) explain how social norms help people judge the appropriateness of their own actions by comparing them to what is typical or what is expected in a specific context. Actions that do not comply with these unwritten social norms can be seen as signs of norm-breaking behaviour. Cialdini et al. (1991) define two types of social norms, descriptive (what one perceives other people to do) and injunctive (What one perceives to be socially acceptable and unacceptable). When applying this to online spaces like comment sections, norm-breaking behaviour like hostility might increase since the user doesn't often face direct consequences. In order to identify norm-breaking behaviour, it is vital to first understand the norms of a given platform.

Differences Between Platforms

Different social media platforms will foster different environments with different social norms. What is considered socially acceptable on TikTok might not be appropriate to comment on Facebook, for example. Graf et al. (2017) found that the perception of what is appropriate in online comments is influenced by the culture of a platform, the level of

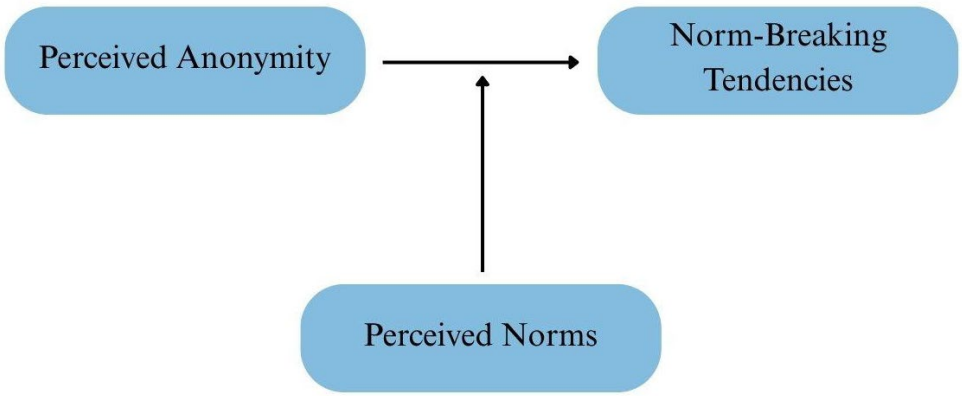
moderation on the platform and the level of identifying factors (perceived anonymity). Levels of anonymity differ across platforms. On Reddit, for example, it is seen as completely normal to have no real name or picture tied to your account. The platform is known for being anonymous, which creates an environment for people to openly share their thoughts and opinions, both harmful and helpful. The same level of anonymity would be considered weird on a platform like Instagram or Facebook, where most people have their full name visible and a clear profile picture of their face. The type of content that gets posted on these platforms also plays a role in what kind of communities the platform fosters. Understanding the differences between platforms is therefore crucial to understanding when behaviour is considered 'norm-breaking'. Users will change their behaviour not only based on their level of perceived anonymity but also on the unwritten (or written) rules of the platform they are on. Lapinski and Rimal (2005) highlight a difference between perceived norms and collective norms and how both can be either injunctive or deductive. Perceived norms are the norms that an individual creates based on the attitude and behaviour they perceive in a given situation and environment. Collective norms are norms that come to exist because of different members of a group or community interacting with each other (Bettenhausen & Murnighan, 1985). Collective norms can only be observed from a social level since asking the individuals in a group about their norms would lead to information on the perceived norm. Even observing collective norms on a social level will lead to perceived norms to a certain extent, since something is always being perceived by someone. Collective norms could therefore also be referred to as 'perceived collective norms', which can explain how social norms and normative beliefs can vary between different social contexts. In order to study norm-breaking behaviour on any social media platform or specific online community, it is important to first gather as much information on the perceived collective norm on that platform. This can be

done by looking at the community guidelines and observing the most common way for people within that community to behave and interact with each other.

With all the definitions of the terms used in our research question now defined, it is possible to create a conceptual model to visualise the variables and their relationship to each other. This visualisation can be seen below in Figure 1. As stated before, this research aims to study the effect of perceived anonymity on a person's norm-breaking tendencies and what effect the perceived norms have on that relationship.

Figure 1

Conceptual Model



Methodology

Experimental Design

To better understand the effect of perceived anonymity on the likelihood of a person's norm-breaking behaviour, this study made use of a 4 condition, between-subjects experimental design. By using this design, the study could examine both the individual effects of the variables included, but also if there is an interaction effect between the two. Participants were randomly assigned to experience one of four conditions. A simple overview of the conditions can be seen in Table 1. Each condition has been given a name for easier results discussion later on. These names are created by taking the tone positive (Pos) or negative (Neg) and the anonymity (Anon) or visibility (Vis) and combining the two into a special code name for each condition. These conditions were set up to test the interaction effect between anonymity and the perceived norms on the platform, as previously seen in Figure 1. The content shown to the participants was consistent across all conditions, and all fell under the category of entertainment.

Table 1

The Four Conditions

Condition	Anonymity	Tone	Name
1	No	Positive	PosVis
2	No	Negative	NegVis
3	Yes	Positive	PosAnon

4	Yes	Negative	NegAnon

Participants

The participants of this study consisted of randomly selected students in the city of Enschede and students who signed up through the SONA system used by the University of Twente, where students can earn SONA credits. A total of 83 participants were recruited, who were all at least 18 years old. All participants were asked to sign a consent form before taking part in the study (see Appendix A). 43 participants identified as male, 37 as female and 3 as non-binary. The average age amongst them was 21.

Procedure

First, the participants were told they were participating in a 'social media behaviour' study and were asked to sign the consent form. This is done to prevent participants from paying too much attention to how 'norm-breaking' their own behaviour may be until the end of the experiment. This experiment makes use of a custom-made, fake social media website that intends to imitate an existing platform; in this case, that platform was TikTok. This website was made by a third-party software developer and is hosted in Firebase. Since the existing social norms and community guidelines of the chosen platform can be perceived, it is also possible to better understand when the norms of this specific platform are being broken. Participants were shown four posts on the TikTok imitation website. Each post had a fake comment section next to it, which had either generally positive or negative comments already posted in it. This showed the participant the tone and norms of this particular commenting environment. Apart from the tone, each fake comment also showed the anonymous status of this website. In conditions one and two, the fake comments are posted by fake users that look real with fabricated first names and last names, using profile pictures that show a real face.

The faces for these profile pictures were gathered from thispersondoesnotexist.com, and the names are randomly generated in order to look real but not use any personal data from real people. In conditions three and four, the fake comments are posted by anonymous users with simple usernames such as ‘user456’ and no profile picture. The participants were informed that this was a simulated environment and were then asked to write one comment on every post they were shown. This procedure took between 5 to 10 minutes per person. Participants were required to comment on each post they were shown to ensure that there was a sufficient amount of data for analysis. This restriction does not accurately reflect a person’s normal commenting behaviour. A few examples of the fake social media page and the comments from different conditions can be seen in Figure 2 and Figure 3 below. Figure 2 shows the positive comment section with non-anonymous users, while Figure 3 shows a negative comment section with anonymous users.

Figure 2

PosVis Example page

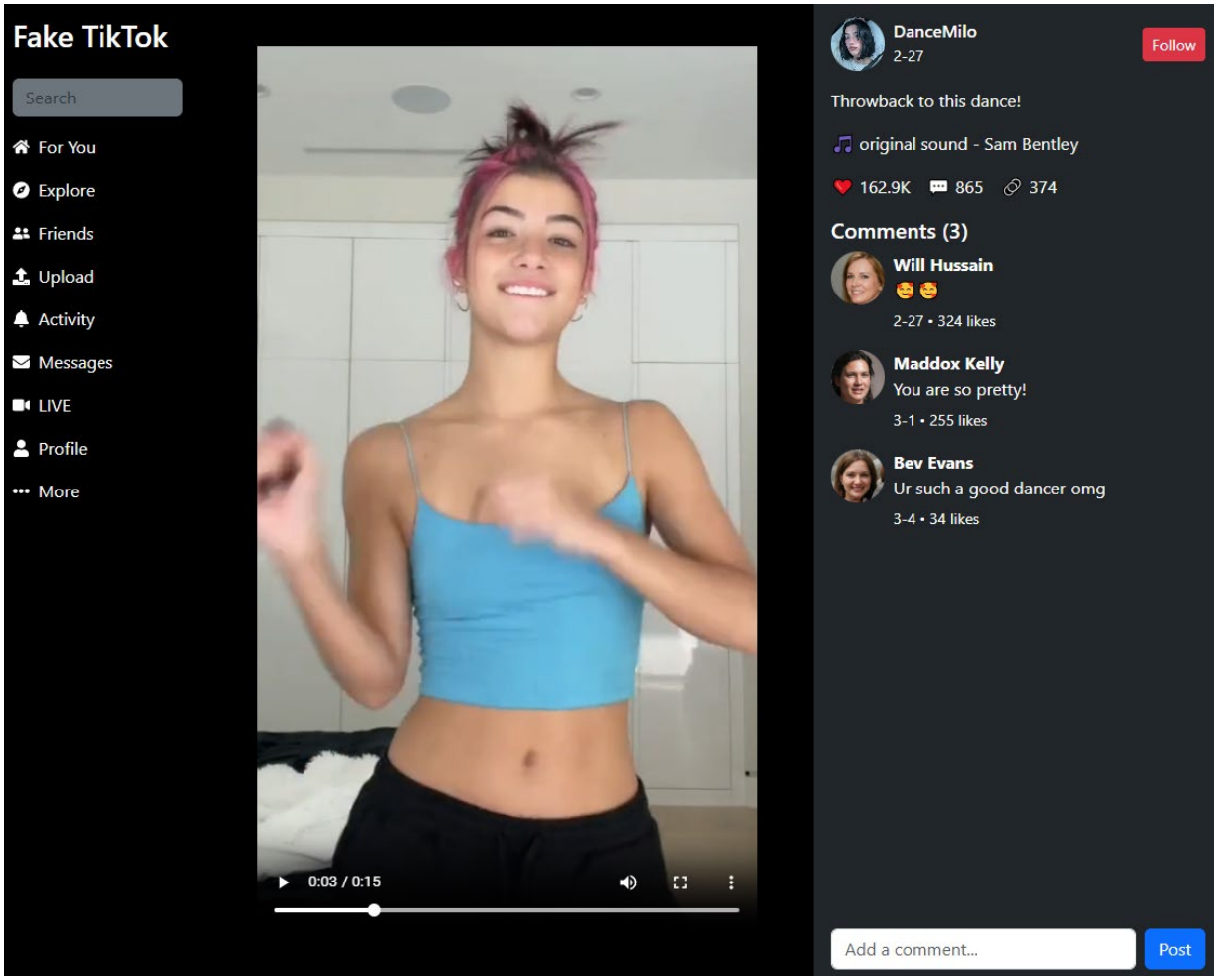
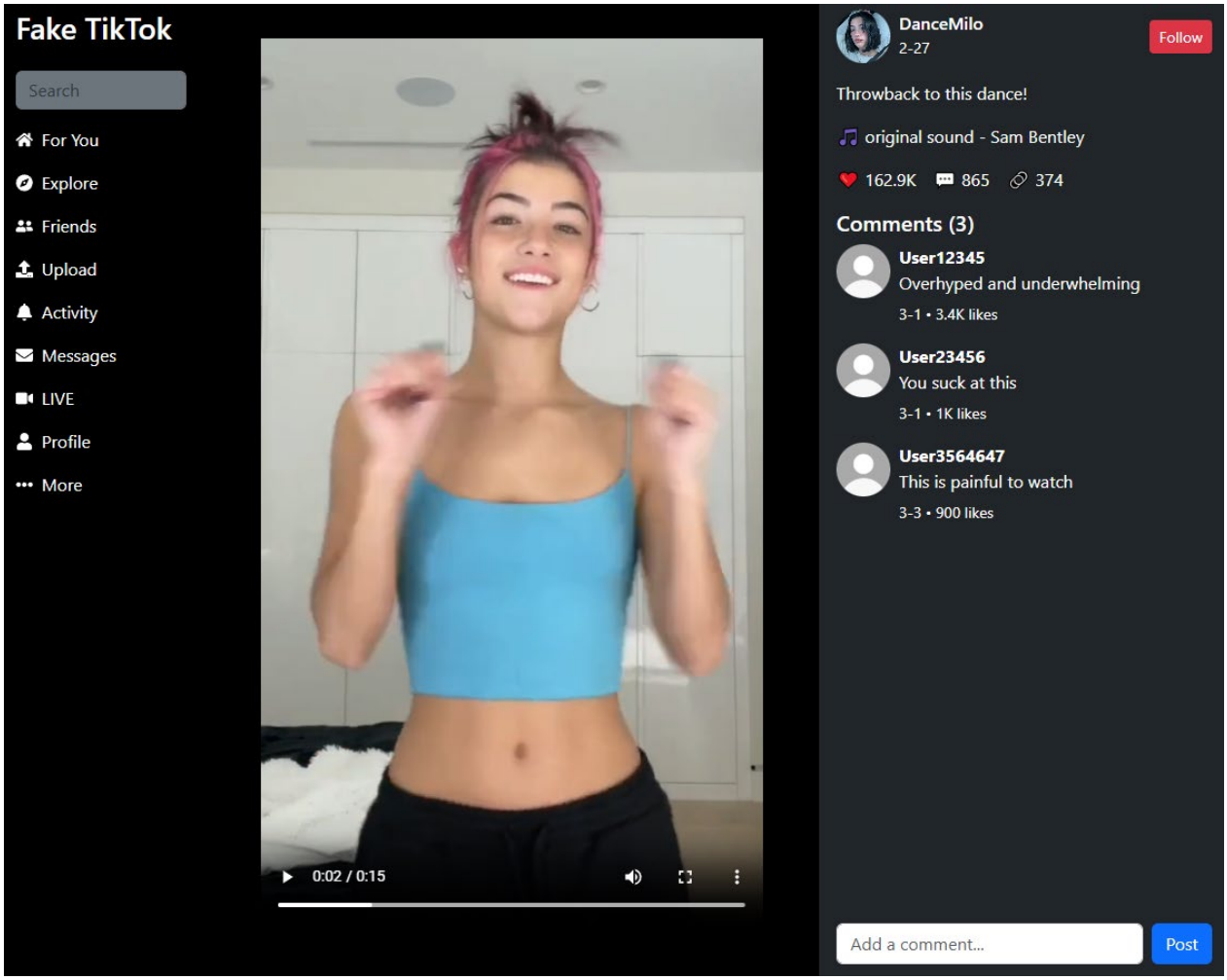


Figure 3

NegAnon Example Page



After the experiment, participants were asked to fill out a short survey regarding their normal commenting behaviour and were informed of the real purpose of the study. Once aware of all the study details, participants were asked if they would like to revoke their consent. The survey included demographic questions about their age, gender, social media use, and likelihood to comment. The participants were also asked how acceptable they find certain norm-breaking behaviours (profanity, insults, community guidelines breaches, etc) using a seven-point Likert scale where 1 = Strongly Disagree and 7 = Strongly Agree.

Analysis

To analyse the content of all the comments in different conditions, each comment was coded to show the level of norm-breaking behaviour shown. What is considered 'norm-breaking' was derived from a combination of community guidelines, the perceived tone and

norms of the fake comment section (positive or negative) and the responses regarding norm-breaking behaviours in the survey.

A coding scheme has been developed based on previous research using a five-point Likert scale, which can be seen in the tables below. Two coders who did not know which conditions each participant was assigned to coded the comments using this scale. To ensure inter-coder reliability was then calculated using Cohen's Kappa. Each comment got an overall norm-breaking score, which was used to generate the mean norm-breaking score per condition. To compare norm-breaking behaviour between all four conditions, a one-way ANOVA test was also done with post-hoc tests when needed.

Table 2

Profanity Coding Scheme

Score	**Description**	**Example**
0	No profanity or offensive terms	"This is alright."
1	Mildly offensive terms	"That's dumb."
2	Non-targeted profanity	"This is bullshit."
3	Targeted profanity	"You look like shit."
4	Repeated or extreme profanity	"Fuck this, fuck you."

Table 3

Hostility Coding Scheme

Score	**Description**	**Example**
0	Calm or neutral tone	"Interesting"
1	Slight condescension	"That's a bit weird."

2	Angry tone or passive aggression	"This is the worst thing I've seen."
3	Openly aggressive/combative	Get your head checked."
4	Threatening, violent, or abusive	"You should be punched for this."

Table 4

Inappropriateness Coding Scheme

Score	**Description**	**Example**
0	Appropriate or neutral	"That's nice."
1	Slight insensitivity	"It's not that deep, get over it."
2	Inconsiderate tone	"You suck, never do this again."
3	Offensive joke	"Lol, you deserve to be bullied for this."
4	Severe violation of moral decency	"I hope you die."

Table 5

Violations Coding Scheme

Score	**Description**	**Example**
0	No clear rule violation	x
1	Mild violation	Spam comments
2	Clear personal insult, mild harassment	"You're useless at this."

3	Targeted hate speech	“You people are the problem.”
4	Harassment, threats, illegal content	Doxxing, threats, etc.

Results

Descriptive Statistics

In this study, a total of 83 participants completed the experiment. Since every participant had to leave 4 comments, there are 332 comments in total. Each participant was randomly assigned to one of the four experimental conditions (Condition 1: n = 22, Condition 2: n = 20, Condition 3: n =21, Condition 4: n = 20). 51.8% of participants identified as male, 44.6% of participants identified as female and 3.6% of participants identified as non-binary. The average age of participants was 21 years old (SD = 1.86) with a range of 18 to 30. Participants were recruited in the city of Enschede and through the University of Twente's SONA system. The survey shows that 86.7% of participants use social media once a day or more, with 56.6% reporting that they check social media multiple times a day. When asked how likely it would be that the participant would comment on the videos seen in the experiment, 83.1% of them answered that it is highly unlikely they would have commented if they had seen these videos in a normal day-to-day setting. 41% of participants also answered that it is highly unlikely that they would comment on TikTok videos in general, no matter the genre. An overview of all descriptive statistics can be seen below in Table 6.

Table 6

Descriptive Statistics Overview

	Amount of Women	**Amount of Women**	**Amount of Non-Binary**	**Mean Age**	**Mean Likelihood to comment (Scale 1 to 7)**
Condition 1 (PosVis)	12	10	0	20.6	2.4
Condition 2 (NegVis)	8	12	0	21.5	1.9
Condition 3	9	11	1	21.9	1.9

(PosAnon)					
Condition 4 (NegAnon)	8	10	2	21.4	2.4

Norms

Participants were asked how much they agreed with certain statements regarding commenting behaviour. The answers are given using a 7-point Likert scale. Each question in the survey is structured in an "It is acceptable to…" followed by an example of norm-breaking behaviour, such as "...use profanity in online comments" or "...insult someone's appearance online.". This means that a 7 means the participants agree with the statement and thus the norm-breaking behaviour, and a 1 means they do not. The answers to these statements can help to gain a better understanding of the general perceived norm of social media users in this experiment. The majority of participants, 83.1% (M = 2.4, SD = 2.2), disagreed with statements that indicate that it is okay to break community guidelines or insult someone's appearance or intelligence. Participants were slightly more lenient when it came to using swear words in comments, with 33.7% (M = 4.2, SD = 2.0) agreeing that they are okay to use. The participants are divided when it comes to joking about sensitive topics, with 44.6% (M = 4.1, SD = 2.1) indicating they do not agree with it. Using sarcasm to criticise others online was seen as acceptable by 59% (M = 4.7, SD = 2) of participants. These numbers give insight into what these participants consider to be acceptable behaviour in online spaces. The differences in norms between conditions were also looked at. The results of this can be seen in the tables below.

Table 7

Condition 1 (PosVis) Norms Overview

	Mean	Median	SD

Profanity Acceptance	3.7	3	2.2
Insensitivity Acceptance	4.7	4	2.1
Insult Acceptance	2.5	2	2.4
Violation Acceptance	2.6	2	2

Table 8

Condition 2 (NegVis) Norms Overview

	Mean	Median	SD
Profanity Acceptance	3.9	4	1.6
Insensitivity Acceptance	4.2	4	2.5
Insult Acceptance	3.1	2	2.6
Violation Acceptance	2.2	1	2.2

Table 9

Condition 3 (PosAnon) Norms Overview

	Mean	Median	SD
Profanity Acceptance	4.7	5	1.9
Insensitivity Acceptance	4	4	1.7
Insult Acceptance	2.4	2	2.4
Violation Acceptance	2.4	1	2.4

Table 10

Condition 4 (NegAnon) Norms Overview

	Mean	Median	SD

Profanity Acceptance	4.9	5	1.9
Insensitivity Acceptance	3.3	3	2
Insult Acceptance	3.4	2	2.7
Violation Acceptance	2.5	2	2.2

Comment Results

Each comment posted was coded by two independent coders in four categories (Profanity, Hostility, Inappropriateness and Violations). Each category was coded on a scale from 0 to 4. The first Cohen's Kappa that was calculated resulted in a 0.57, which indicates a moderate agreement between coders. After reviewing the codebook and making adjustments to definitions using clarifying examples, the data were recoded. After recoding the data, a new Cohen's Kappa of 0.96 was calculated, which shows an extreme agreement. An overview of the average scores of each condition can be seen below in Table 3.

Table 11

Comment Mean Scores

	Profanity	Hostility	Inappropriateness	Violations
Condition 1 (PosVis)	0.017	0.189	0.143	0.011
Condition 2 (NegVis)	0.025	0.379	0.379	0.063
Condition 3 (PosAnon)	0.022	0.131	0.131	0.012
Condition 4 (NegAnon)	0.075	0.325	0.3625	0.05

ANOVA

Apart from the comments' mean scores, a one-way ANOVA test was done to compare the levels of profanity, hostility, inappropriateness and community guideline violations. The analysis showed that the condition that the participants were in did not have a significant effect on the amount of profanity used ($F(3, 330) = 2.19$, $p = .089$, $\eta^2 = .019$). Post hoc comparisons using Tukey's HSD test did not show any significant differences between conditions (all $p > .11$). The amount of hostility does differ significantly across conditions ($F(3, 330) = 3.54$, $p < .015$, $\eta^2 = 0.031$), with the post hoc comparison showing that condition 2 had significantly higher hostility scores compared to other conditions (M difference = -0.25, p=.025). Another significant effect was seen for inappropriateness ($F(3, 330) = 5.37$, $p = .001$, $\eta^2 = .05$). Tukey post hoc comparisons revealed that Condition 2 (M difference = 0.24) and Condition 4 (M difference = 0.22) both had significantly higher Inappropriateness scores than Condition 1 (both p<.05p < .05p<.05). No significant differences were found across conditions for violations ($F(3, 330) = 1.35$, $p = .259$, $\eta^2 = .01$). Post hoc comparisons confirmed that there is no significant differences between conditions (all $p > .36$).

Discussion

This study aimed to gain a better understanding of the effects of perceived anonymity on the likelihood of norm-breaking behaviour occurring in online comment sections. The comments posted in the four different simulated environments were coded to find their levels of profanity, hostility, inappropriateness and community guidelines violations.

Interpretation

The results of this study show that the condition the participants were in did significantly influence the overall 'norm-breaking' levels of the comments posted. Conditions 2 and 4 (NegVis and NegAnon), which were the conditions with negative comments next to the videos, showed significantly higher hostility and inappropriateness scores than the conditions with positive comments. This suggests that showing a more norm-breaking behaviour from others, whether they are anonymous or not, has an effect on how likely participants were to engage in norm-breaking behaviour themselves, which falls in line with the study from Kahneman and Miller (1986). There were no significant differences between conditions when it came to community guidelines violations, which is in line with the results of the survey, where the majority of participants indicated that breaking guidelines was not okay.

The survey also gave insight into the general norms and behaviour that participants deemed acceptable. Although the majority disagreed with statements that said it was okay to leave hateful or insulting comments, there was more tolerance towards behaviour like sarcasm, swearing or joking about sensitive topics. This shows that these types of behaviour are likely more normalised in online spaces such as TikTok, where 'norm-breaking' humour and casual, informal language are more common. This is in line with the study done by Graf et al. in 2017, where they explain how different levels of moderation and the different

cultures of online platforms cause people to behave very differently depending on which social media sites they use.

Many participants pointed out that they do not post comments on TikTok videos in their day-to-day lives, despite being on social media quite often. The overwhelming majority of participants said they would have never commented on the videos that were shown in this experiment, despite the original videos being extremely popular on TikTok. The likelihood of commenting does not differ much between conditions. It is equally unlikely in all.

Implications

The differences in norm-breaking behaviour between the 'positive' conditions 1 and 3 and the 'negative' conditions 2 and 4 show that social media users can fall into a kind of feedback loop. Just a few negative comments can lead people to follow by example and mirror the behaviour of others. This means that the norms within a specific comment section can change depending on the first comments people see and accept as the norm, leading to more negative comments. This can become a risk on platforms like TikTok or YouTube, where comments are easily visible and moderation tends to be lacking. Platforms like these should not treat comment sections as something neutral but rather as an environment that shapes and affects people's behaviour. Moderating comments more actively can help prevent users from falling into this feedback loop of negativity and create a more civil space for people to comment and discuss content. This moderation can be done by the platform but also by the content creators themselves.

This study also shows that norm-breaking behaviour increases even if the user is made to feel non-anonymous. Participants did not seem to care that the other comments seemed to be placed by real people, whose faces were clearly on display right next to their full names. Norm-breaking behaviour increased without high perceived anonymity. This would suggest that perceived norms are a more powerful influence than just anonymity alone.

This would indicate that more active moderation would have a greater effect on the amount of norm-breaking behaviour than forcing a level of identifying characteristics on someone's profile. It also shows that negativity in comment sections will likely draw in more negativity, but this also works the other way around. Positive comments are likely to provoke negativity and thus prevent negativity from spreading. This means that although moderation helps, actively combating negative comments by liking and uplifting positive ones would also make a difference to the perceived norms in comment sections.

Limitations and Future Research

A major limitation of this study is the limited number of participant responses acquired for analysis. A larger sample of at least 30 participants in each condition would allow for more reliable conclusions and an increased level of generalizability to a broader population of social media users. While some significant effects were found, they should be examined with caution, as only a replication study with a larger sample size would be able to confirm their reliability.

In addition to the small sample size, this study was conducted in person and on the researcher's laptop, with the researcher present while participants filled in their answers. While this did ensure that participants finished their tasks and had guidance while navigating the fake, simulated environment, it may have also affected how participants responded. Participants might have felt influenced or intimidated by the presence of the researcher and changed their commenting behaviour to seem more generally socially acceptable or nice. This would affect the norm-breaking scoring done by the coders later on, since participants might have felt watched or even judged despite being instructed to comment however they felt like.

Lastly, this in-person setup does not accurately represent a person's normal environment when engaging with content on social media. In real life, users are often alone,

at home and use their own devices. These differences might have also affected their sense of anonymity and comfort, which might affect how accurately this study reflects real-world commenting behaviour. Many participants said something about how they "Would never comment on these things in real life" and how they "don't know what to say." which shows how unnatural the experiment felt to them. Future studies could benefit from creating a website that is easy to use on smartphones, that participants can access and understand without help from the researcher. This would allow them to go through the steps of the experiment in a comfortable environment where they are more likely to act as they would normally.

Conclusion

This study aimed to understand the extent to which anonymity affects norm-breaking in online comment sections and how the perceived norms of a comment section might influence that relationship. Building off of previous research about Norm theory and the Online Disinhibition effect, this research aimed to fill the gap in research when it comes to perceived anonymity, specifically, since almost no one online is completely anonymous. While people might assume that anonymity causes norm-breaking behaviour to spike, this study shows a slightly more nuanced picture. The results show that the tone of a comment section plays a larger role in people's likelihood to be hostile or inappropriate. People feed into negativity and tend to copy what they see, no matter how seen or invisible they feel. This shows that the perceived norms within an online comment section might overpower the effect of anonymity alone.

Now, imagine encountering hateful comments under something like a post about the Barbie movie or a random TikTok trend. It seems that the people in the comments have fallen into a negative feedback loop where just a few hostile words have caused the tone of this comment section to feel negative. This is not mindless outrage; these people are shaped by

their online environment, which is filled with other people showing them what to do. One mean comment can change the norm, causing more to follow. By understanding the possible cause behind the negativity, it is possible to start fostering a more respectful and kind online environment. As this study shows, negative comments attract slightly more negativity, but positive comments also attract positivity. Recognising the power of perceived norms in these online spaces is an important step towards the creation of online spaces that can encourage both freedom of expression and respectful interactions.

References

Bettenhausen, K., & Murnighan, J. K. (1985). The emergence of norms in competitive Decision-Making groups. Administrative Science Quarterly, 30(3), 350. https://doi.org/10.2307/2392667

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A Focus Theory of Normative Conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology* (pp. 201–234). https://doi.org/10.1016/s0065-2601(08)60330-5

Graf, J., Erba, J., & Harn, R. (2017). The role of civility and anonymity on perceptions of online comments. *Mass Communication & Society*, *20*(4), 526–549. https://doi.org/10.1080/15205436.2016.1274763

Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods Instruments &Amp Computers*, *31*(3), 433–438. https://doi.org/10.3758/bf03200723

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153. https://doi.org/10.1037/0033-295x.93.2.136

Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication Theory*, *15*(2), 127–147. https://doi.org/10.1111/j.1468-2885.2005.tb00329.x

Marx, G. T. (1999). What's in a name? Some reflections on the sociology of anonymity. *The Information Society*, *15*(2), 99–112. https://doi.org/10.1080/019722499128565

Rimal, R. N., & Lapinski, M. K. (2015). A Re-Explication of social norms, ten years later. *Communication Theory*, *25*(4), 393–409. https://doi.org/10.1111/comt.12080

Scott. (2004). Benefits and drawbacks of anonymous online communication: legal challenges and communicative recommendations. *Free Speech Yearbook*, *41*(1), 127–141. https://doi.org/10.1080/08997225.2004.10556309

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Appendix A

Participant Consent Form

Social Media Behaviour Study

Dear participant, You are invited to participate in a research study about social media behaviour. This study is being conducted for a bachelor's thesis in Communication Science at the University of Twente.

We are interested in the commenting behaviour of students on social media. If you agree to participate, you will be asked to leave a few comments on some videos on a fake website and fill in a short survey. It will take you approximately 5 to 10 minutes to complete.

Please know that your participation is entirely voluntary and that you can stop at any point, for any reason, without consequences. All responses will be anonymous and are only visible to the researcher and supervisor. The data will only be used for academic purposes and will be deleted after this study is completed.

If you have any questions or concerns, feel free to contact the researcher at s.bernardus@student.utwente.nl

- I am 18+
- I have read and understood the information above
- I agree to take part in this study

Appendix B

Use of AI Disclosure

During the preparation of this work, I used GPT-4o-mini in order to generate the code needed to analyse the data gathered from participants. After using this tool, I reviewed and edited the content as needed, and I take full responsibility for the content of the work.

Appendix C

Literature Log

The table below shows a clear overview of the sources used in this paper, where they were found, which keywords were used to find them and why I decided to keep them. Some sources were also found through the reference list of other sources or recommendations from peers. Unfortunately, it is not currently possible to trace back the steps that were taken when the sources for this research were selected. This log is a recreation of what the log should look like, but does not contain information on sources that were not chosen or exact word strings that were used.

Table 12

Literature Log

No.	Source	Database	Keywords	Reason for Keeping
1	Bettenhausen, K., & Murnighan, J. K. (1985). The emergence of norms in competitive decision-making groups. *Administrative Science Quarterly, 30*(3), 350. https://doi.org/10.2307/2392667	Web of Science	Norms, Behaviour, Social	Shows how behaviour changes in groups, which is relevant to online behaviour.
2	Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct... *Advances in Experimental Social Psychology* (pp. 201–234). https://doi.org/10.1016/s0065-2601(08)60330-5	Scopus	Social, Norms, Behaviour	Helps define norms.
3	Graf, J., Erba, J., & Harn, R. (2017). The role of civility and anonymity on perceptions of online comments. *Mass Communication & Society, 20*(4), 526–549. https://doi.org/10.1080/15205436.2016.1274763	Web of Science	Anonymous, Anonymity, Online, Commenting, Behaviour	Recent study on how anonymity affects how people behave online.
4	Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods,*	Google Scholar	Anonimity ,Behaviour	Discusses how anonymity affects self-reporting online.

	Instruments, & Computers, 31(3), 433–438. https://doi.org/10.3758/bf03200723			
5	Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136–153. https://doi.org/10.1037/0033-295x.93.2.136	Scopus	Norms, Behaviour	Helps understand norms and how they shape behaviour.
6	Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication Theory, 15*(2), 127–147. https://doi.org/10.1111/j.1468-2885.2005.tb00329.x	Web of Science	Norms, Social, Behaviour	Helps define norms.
7	Marx, G. T. (1999). What's in a name? Some reflections on the sociology of anonymity. *The Information Society, 15*(2), 99–112. https://doi.org/10.1080/019722499128565	Google Scholar	Anonymity, Online	Explores the sociological aspect of anonymity, which helps understand online anonymity in modern times.
8	Rimal, R. N., & Lapinski, M. K. (2015). A re-explication of social norms, ten years later. *Communication Theory, 25*(4), 393–409. https://doi.org/10.1111/comt.12080	Scopus	Norms, Social, Online	An updated theory on social norms which supports current research.
9	Scott. (2004). Benefits and drawbacks of anonymous online communication: Legal challenges and communicative recommendations. *Free Speech Yearbook, 41*(1), 127–141. https://doi.org/10.1080/08997225.2004.10556309	Google Scholar	Anonymity, Social, Media, Commenting, Anonymous	Discusses the pros and cons of anonymity online.
10	Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*(3), 321–326. https://doi.org/10.1089/1094931041291295	Google Scholar	Anonymity, Behaviour, Social Media	Explains why people might behave differently online. Important when looking at anonymity and commenting behaviour.