THE UNIVERSITY
of EDINBURGH

Expression and perception of identity through skin-toned emoji

Alexander Robertson

Doctor of Philosophy

Centre for Doctoral Training in Data Science

School of Informatics

University of Edinburgh

2022

Abstract

The introduction of emoji skin tone modifiers to the Unicode Standard in 2015 was met with considerable debate on the extent to which these emoji would be used, who would actually use them, and what they would actually be used for. I evaluate such claims against large datasets drawn from social media and find evidence that people generally *produce* skin-toned emoji which align with their real-world identity. I also identify particular variations in emoji production based on real-life skin tone and geographical location, as well as the context in which emoji are used. I test experimentally the extent to which people *perceive* identity from the emoji that others produce, finding that these emoji are strongly considered to represent specific identities for both authors and readers of social media. Even default yellow emoji without a skin tone are associated with a particular identity. The ability of emoji to index identity for both author and reader is a property found in language, where one consequence of this is a difference in attitudes, responses or behaviours towards perceived identities. Whether the indexicality of emoji can similarly affect *behavioural outcomes* is tested experimentally, where no such effect is found.

Acknowledgements

Foremost, there are the people without whom this particular thesis would not have happened. Sharon Goldwater and Walid Magdy have been a supervisorial dream team. Any problems I had over the last five years were solely attributable to either me or a global pandemic --- so few people undertaking a PhD can make such a claim and I will be forever grateful that I can always look back on this time fondly. The details are fuzzy now, but it was Walid who suggested looking at emoji and I am glad we did and that Sharon approached it with the seriousness she does with any research topic.

It was Lauren Hall-Lew who introduced me to the experimental methodology used in Chapter 4 as well as the wider sociolinguisic material that this thesis draws upon. Lauren was also the first to politely laugh at the timeline I proposed during my first year review, and was right to do so. But everything got done in the end, even if it took *slightly* longer than the eight months I predicted. I wonder what I was planning to do with the remaining year and a half of funding?

Chris Lucas kindly supported me with the statistical analyses in Chapters 4 and 5 and reassured my supervisors that I wasn't just making stuff up as I went along. Perhaps just as important were our trips to the pub with Kate.

In addition to teaching me a lot about designing and running experiments, Hugh Rabagliati introduced me to preregistration when I was his research assistant in 2015. If I could offer you only one tip for the future, preregistration would be it. I cannot overstate how important this was to making my life easier, as well as adding a conspicuous aspect of rigour to my work. You would be surprised how easy it is to write a paper when armed with a document you wrote months before, listing what data you will collect, what you predict, and how you will analyse the data.

Then, there are the people without whom I may not have ever made it to the point of writing any thesis. Paula Buttery not only admitted me to Cambridge in

the first place but introduced me to computational linguistics and supervised my BA dissertation. Bettina Beinhoff, in addition to being an excellent Director of Studies with great taste in whiskey and beer, introduced me to sociolinguistics. Their support, guidance and confidence in me then still helps me keep my chin up today.

Chris Latimer has been a constant source of strength and support, either quietly leading by example or more overtly helping when life was difficult, over the last twenty or so years. If a PhD thesis on skin-toned emoji usage is in any way capable of repaying even the smallest part of his love and kindness, then I gladly dedicate this thesis to him.

Finally, there are those who are just a joy to be around and made PhD life interesting and fun. Anny, who I began my acacemic journey with all those years ago at Ruskin and has always had my back. Viktorija, the best thing to come out of that LSA summer school. My CDT peers: James, Kate, Maria, Matt, and Pippa. My non-CDT peers: Janie, Deena and Lewis. Fellow survivors of that Bell Labs internship: Katrin, Luis, Matheus, Melanie, Milan, Xi and Yongsung. Michela, who after leaving Edinburgh was conveniently in Cambridge every summer I was interning there and always up for a beer in the sun. Oliver, for being an excellent housemate before and during the pandemic. Dmitriy, for also being a great housemate when I first moved to Edinburgh and for introducing me to Andrew who never became visibly annoyed at being repeatedly asked ‘‘What plant is this?’’ when out on walks. Friends from Wolfson who have cheered me on since 2012: Anders, Anne, Ben, Brian, Edward and Ryan. And finally my beloved lingos: Amy, Emma, Immy, Jess and Sandy ❤️

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Alexander Robertson*)

Table of Contents

Chapter 1

Introduction

Identity, and how we express it, is at the core of the human social experience. The language we use, and the way we use language, is intricately connected to our identity. The linguistic properties of our communications mark us as a particular person, one with particular characteristics. We do this for any number of reasons, in any number of ways. I might roll my r's more than is strictly necessary in order to ensure someone knows I am Scottish. You might use a specific word for a bread roll to mark yourself as being from a specific region of England. Yet another person might specifically avoid that word, in an effort to hide being associated with that region. Such linguistic performances are deliberate and exploit a shared understanding between speaker and listener: that specific facets of identity are indexed by specific linguistic features. Such performances are not always deliberate and the finer linguistic details of speech are generally automatic and subconscious, but the understanding of which aspects of identity are linked with which linguistic properties is always in play at some level of awareness for a listener.

Which performances are associated with which identities has been the subject of decades of sociolinguistic research. Common linguistic features are variations in syntactic structure and in phonetic realisation. These have been linked to a wide

variety of identity aspects such as social group membership and more general demographic details such as age and gender. Experimental methodologies have mainly employed recordings of speech, as linguistic variation is most amenable to precise manipulation for experimental conditions. However, not all communication, and therefore not all expressions of identity, takes place through speech. Computer-mediated communication, especially social media platforms with human connection and interaction as core services, offers new affordances for us to let others know who we are and what we are about. One such novelty is emoji, a constrained set of icons which can be used alongside text and have become particularly commonplace in online communications over the last decade. Since 2015, users have been able to apply one of five different skin tones to a subset of emoji that represent human faces, bodies and body parts.

As a starting point I investigate patterns usage in this subset of emoji and evaluate specific claims made (but not always tested) by the media regarding who, if anyone, would use such emoji and why. The next question of this thesis is whether emoji used on social media can communicate identity and social meaning between authors and readers. This is motivated by my observation that skin-toned emoji are almost exclusively used in a self-representational manner. Finding that emoji do communicate social meaning, the final question is whether which identity is communicated can influence the behaviour of others.

1.1 Data: Production and Usage

The first component of this thesis is a large scale analysis of data drawn from Twitter linked with crowd-sourced annotations of user demographics. This provides descriptive statistics on the usage of tone-modifiable emoji, addressing a range of exploratory questions based on both unsubstantiated claims made elsewhere and patterns observed in the data.

The main contribution of this component is a quantitative analysis of how skin-toned emoji are used on social media, which I present within the context of sociological frameworks for self-representation and identity. This replaces previous unsubstantiated claims and assumptions on the matter: these emoji are widely used by users of all skin tones; racist usage is not corroborated; usage is overwhelmingly self-representational. I also explore the phenomenon of users selectively applied skin-toned emoji and provide a qualitative analysis of the situations in which this can occur. Finally, I offer a design-based account of why we observe lower rates of use of the darkest emoji skin tone.

1.2 Experiments: Perception and Behaviour

The second component is a series of carefully controlled pre-registered experiments, drawing on established methodologies in sociolinguistics.

The first builds on the observation from data analysis that people do indeed use skin-toned emoji which reflect their physical appearance. It measures the extent to which readers associate the tone-modifiable emoji they see with authors of specific ethnic or racial identities. In the second, I test whether knowledge about an author's identity, as derived from their emoji usage, can influence how others behave towards that author's communications.

1.3 Thesis outline

Overall this thesis connects the emoji that people *produce* with what aspects of identity people *perceive* from such usage, then connects this to how people *behave* with the information about identity they glean from emoji.

Chapter 2: Background This thesis draws upon theoretical and experimental works in linguistics and sociology, focusing on a technical aspect of computer-

mediated communication --- the background chapter is necessarily broad. I provide an historical and technical overview of emoji and their relation to the Unicode Standard, introducing the terms of art used throughout this work. The concepts of identity and self-representation, central to this thesis, are defined with reference to major works in sociology. I then review works from sociolinguistics on the connection between language and identity.

Chapter 3: Production This chapter quantifies the extent of and variation in usage of tone-modifiable emoji --- their production. The first part describes the data collection process and analyses performed, answering the questions listed above. The second part addresses specific phenomena observed in the first. I show that tone-modifiable emoji are widely used and in a manner that links them closely to expression of identity.

Chapter 4: Perception Building on the findings of Chapter 3, I conduct experiments designed to test the extent to which readers of tone-modifiable emoji, rather than authors, connect those emoji to specific identities. Results are presented in the context of pre-registered hypotheses, research questions and analysis plans.

Chapter 5: Behaviour The final chapter tests whether emoji, shown in Chapter 4 to be a salient indicator of author identity, can influence the behaviour of those who observe the emoji in communications. Results are presented as in the previous chapter.

Chapter 2

Background

As stated in Section 1.3, this thesis draws upon work from a variety of areas. This background section serves to introduce the key concepts taken from those areas, as well as provide an overview of emoji. The chapters that follow will focus on details of implementation and analysis.

2.1 Emoji and Unicode

The word *emoji* comes from the Japanese phrase 絵文字 (“picture character”). Emoji were first available in 1997, for use on the DP-211SW mobile phone released by the Japanese telecoms company SoftBank (Burge, 2019). In 1999, a larger set of emoji were available to users of the i-mode service created by NTT DoCOMO

Figure 2.1: The 90 emoji available on the DP-211SW mobile phone, in 1997

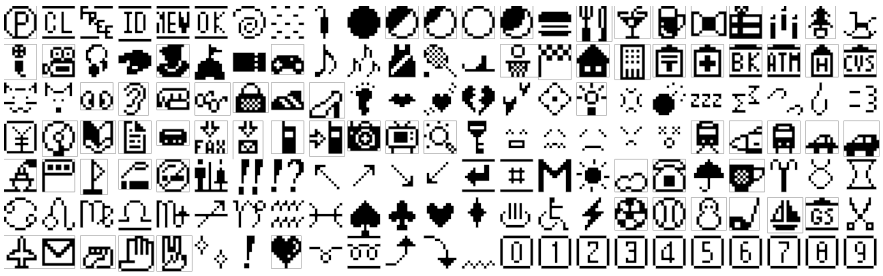


Figure 2.2: The 161 emoji available on i-mode mobile phones, in 1999

(Galloway, 2016). These icons were inspired by pre-existing symbols. Some were specific to Japan, drawing on anime and other cultural aspects, while many of the larger 1999 set will be recognisable to those familiar with Zapf Dingbats or Wingdings --- fonts which provide symbols or ornamentation for use alongside text. All emoji were monochrome and had a height and width of 12 pixels, as shown in Figures 2.1 and 2.2.

The NTT DoCOMO emoji were designed by Shigetaka Kurita, who felt that digital communication in Japanese was hindered by an inability to contextualise the many ritualised and honorific expressions used in traditional written and face-to-face communication. ''If someone says 'Wakarimashita' you don't know whether it's a kind of warm, soft 'I understand' or a 'Yeah, I get it' kind of cool, negative feeling. You don't know what's in the writer's head.'', said Kurita in a 2013 interview (Blagdon, 2013).

Initial emoji offerings were service-specific. An emoji sent by a user of a particular network could only be viewed by a recipient on the same network, using a compatible phone likely provided by the network operator, due to how emoji were encoded by carriers. The Japanese language was encoded using the Shift JIS [1] character set. Such sets map characters of a language to a sequence of one or two bytes, known as codepoints. Not all codepoints were used for standard language characters and so these were used to encode emoji, but there was no standardisation between service providers on which emoji were represented by

[1] `https://en.wikipedia.org/wiki/Shift_JIS`

which byte sequences. This lack of standardisation continued until 2008, when Apple implemented emoji in iPhone OS 2.2 for Japanese users on the SoftBank network. In 2010, these emoji were made available for phones worldwide as part of iOS 4.0 as well as for computers running OS X 10.7.

The inclusion of emoji in Apple systems worldwide placed emoji in the public eye. Their popularity and widespread usage saw other vendors, among them Google and Microsoft, implement emoji for their own products. Each had a distinct visual style, but by this point in time standardisation of the codepoints behind the emoji was agreed upon through the Unicode Consortium. Unicode 6.0 explicitly encoded emoji and assigned codepoints to aid cross-platform compatibility and interoperability.

Two issues arose from including emoji in the Unicode Standard. First was the fact that many characters that were now considered to be emoji were actually already represented in previous Unicode versions. These were mainly Dingbat-type characters (e.g. envelope, telephone, heart) and already had their own codepoints in other sections of the Unicode Standard. This raised the question of what makes a character an emoji. This was driven by the second issue --- increasing demand for new emoji. Many demands were highly specific to particular cultures or groups of people, often seeking parity with existing emoji or to fill in gaps in existing sets of emoji. Berard (2018) describes the process by which emoji are created. From 2010 to 2015, three major versions of the Unicode Standard were released. The number of included emoji grew from 747 (v6.0) to 858 (v7.0) to 1634 (v8.0).

In 2015, the Unicode Consortium consolidated all emoji-related codification in Emoji 1.0. This release clarified which pre-existing codepoints from which previous Unicode releases were to be considered emoji. This distinction was important, ensuring that vendors producing their own visual designs for emoji were apprised of exactly how many icons they should create. Emoji v1.0 essentially decoupled the administration of emoji from the main Unicode Standard, allowing them to be

developed as their own entity. The version naming of this documentation changed in 2017, to match the main Unicode Standard versions. As of 2021, Unicode and Emoji are on v13.0, with v14.0 to be released by the end of the year. That version will add 37 new emoji to the current 3521.

That number is perhaps misleading, due to how characters are represented as codepoints. The Unicode Standard encodes only *characters* and does so by using UTF-8 to define *codepoints*. Unicode does not define *glyphs* --- the final rendered item a user actually sees on a device. UTF-8 is a variable-width system --- a character may be represented using between one and four individual bytes. There are 1,112,064 valid codepoints in UTF-8. One aim of the Unicode Consortium is to standardise the digital representation of all the characters of the world's writing systems, but the supply of codepoints is finite. A single glyph may therefore be composed of a *sequence* of codepoints.

This technical detail takes on a quite literal compositional aspect in emoji. The emoji for scientist 👨‍🔬, rather than having its own codepoint, is instead a sequence of 👨 + `U+200D` + 🔬. Codepoint `U+200D` is the zero-width joiner (ZWJ) and signals to a system parsing UTF-8 encoded text that the adjacent codepoints are to be merged in some way --- this prevents adjacent emoji from being automatically merged. ZWJs are also used to place diacritics and merge characters in writing systems such as Arabic and many Dravidian languages. Therefore, the number of emoji we might count on our keyboard is greater than the number of emoji codepoints found in the Unicode Standard. The use of ZWJs also allows emoji to represent world flags, using a combination of 26 letter codes, and to provide gendered versions of some emoji by using either a man/woman version of the person emoji, or adding the male/female symbol after another ZWJ.

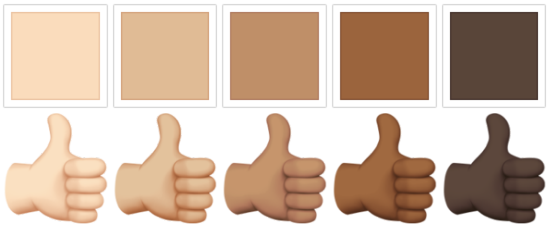


Figure 2.3: The five skin tones (top) and example emoji (bottom) as used by Apple

	Example	Counter-example
TME	🤷‍♀️👋	😍💩
TME+	🧔🏿	👨
TME-	🤷‍♀️	🤷‍♀️

Table 2.1: Examples of the terms TME, TME+ and TME- in use, with counter-examples.

2.2 Emoji and skin tone

In 2015, Unicode v8.0 specified five new emoji codepoints (`U+1F3FB` to `U+1F3FF`) representing skin tones based on the Fitzpatrick Scale (Fitzpatrick, 1988), a dermatological categorisation of skin types based on their reaction to sunlight. Examples are shown in Figure 2.3. These codepoints can *only* be used with a specific subset of humanoid emoji, and therefore do not require a ZWJ. The effect of these codepoints is to change the displayed glyph to one with a specific skin tone. When no such codepoint is present, the glyph is displayed with the bold yellow associated with "smiley" emoji such as 🙂 or 🙄.

Throughout this thesis I will refer to emoji with this specific property as *tone-modifiable emoji*, or TME. When referring to any TME to which a skin tone modifier has, or has not, been attached then I use the terms TME+ and TME- respectively. TME therefore denotes the class of emoji which can be modified in this way, while TME+ and TME- denote specific realisations of TME by users of emoji. See Table 2.1 for specific examples.

Figure 2.4: The 471 emoji available on Japanese iPhones running OS 2.2, in 2008, and worldwide in 2010 with iOS 4.0.

Prior to v8.0, most humanoid emoji were light-skinned by default, as shown in the example of Apple's original emoji in Figure 2.4. This was acknowledged in section 2.2 of Unicode Technical Report #51 (Davis and Edberg, 2014), where the motivation for this change was outlined: ''People all over the world want to have emoji that reflect more human diversity, especially for skin tone. The Unicode emoji characters for people and body parts are meant to be generic, yet following the precedents set by the original Japanese carrier images, they are often shown with a light skin tone instead of a more generic (non-human) appearance, such as a yellow/orange colour or a silhouette.''

That bold, decisive statement belies the tortuous process which led to the addition of TME to the Unicode Standard. Miltner (2020) provides a reconstruction of this process, based on Unicode Consortium mailing lists and an interview with a Consortium member who was heavily involved with emoji changes in the Standard.

Miltner argues that the lack of racial representation prior to Unicode v8.0 was because Consortium members simply refused to acknowledge that there *was* a lack of diversity and were unwilling to discuss the political nature of technology, in particular the tendency to insist that technology is utterly divested from the physical form and its large degree of variation, and is therefore neutral by default.

Media response to TME was mixed. For some, skin tones were not needed and emoji should be ‘‘raceless with yellow faces’’ (Tutt, 2015), especially since the new emoji were just recoloured versions of the original white emoji --- ‘‘bastardized emoji blackface’’ (Tutt, 2015). Others were of the opinion that yellow emoji were naturally neutral, therefore representative of anyone and everyone (Pardes, 2015; Dewey, 2015). The issue of ‘‘digital blackface’’ was also raised by Dickey (2017), drawing on personal examples of people using TME+ that do not match their actual skin tone, while Princewill (2017) highlights the negative use of such TME+ as a form of cultural appropriation. Would TME+ even be used at all? Besides people being too lazy to take the additional steps of holding down an emoji on the keyboard then selecting a skin tone, doing so ‘‘just seems so declarative’’ while the options available, while expanded, were ‘‘not close to my color at all’’ (Zimmerman, 2015). And based on analysis of 18,000 tweets gathered in the US there were claims that white people in particular rarely use tones at all (McGill, 2016).

I group most of these questions of TME usage under the issue of *production* --- how people use emoji in their online communications. In Chapter 3 I show that Twitter users do indeed use TME+. The TME+ produced by Twitter users are a close match to their real-world skin tone and use of non-matching TME+ is rare and not associated with negative acts of ‘‘digital blackface’’. The specific question of whether default emoji are neutral is addressed experimentally in Chapter 4, where I show that they are in fact not neutral and instead are generally perceived as representing a white identity.

2.3 Identity and self-presentation

In Chapter 3 I will argue that the observed production of TME by Twitter users is evidence that TME are self-representational, used to express one's identity. While the Unicode Consortium did not explicitly mention race or ethnicity when it added more options to represent human diversity, skin tone is unarguably one dimension of those particular aspects of human diversity. Identity as understood through skin tone is the focus of the experiments in Chapter 4 and in that chapter I go into more detail on the terminology used and experimental design decisions. Here, I provide an overview of the sociological frameworks I have relied upon to understand identity and self-representation. The former is Jenkins' conception of social identity (Jenkins, 2014), the latter Goffman's approach to self-presentation (Goffman, 1956).

The first of these is Jenkins' work on identity and his argument that identity is a process --- identification. It is not a *thing* that we should reify as an *is* (Jenkins, 2014, p. 17). The identification process uses measures of similarity and difference, to give us our sense of who is who. Identity is not something we *have* but essentially a classification task that we *do*. This classification is multi-dimensional both in the number of identities we may associate with any one person and also the number of cues we use to make those classifications. These cues naturally involve the embodiment of an individual through their physiology, clothing, and adornments such as tattoos and jewellery. Behaviour and mannerisms are also relied upon as identity cues. But the core of identity, and the human capacity for knowing who is who, is ''rooted in language''. Being identified as belonging to any particular class may have consequences, seldom neutral --- Jenkins (2014, pp. 2-5) gives multiple examples of the negative repercussions of identification. Repercussions, negative or otherwise, are not a natural outcome of the identification process. The impact of identity depends on us to enact that impact --- ''identification has to be *made* to matter'' (Jenkins, 2014, p. 6).

These are the two main aspects of identity considered in this thesis. First, whether TME are a cue that can be used by people as part of the identification process. This is the central question of Chapter 4. Second, having determined that they are indeed such a cue, whether this matters to people, in terms of influence on behaviour. In Chapter 5, I present evidence that they do not.

Jenkins' social identity framework, with its focus on identity as a process, is complementary to earlier work by Goffman (Goffman, 1956). Goffman's dramaturgical approach examines face-to-face interaction through the metaphor of the theatre: ''Within the walls of a social establishment we find a team of performers who co-operate to present to an audience a given definition of the situation.'' (Goffman, 1956, p. 152) People are actors, performing different roles in society for different audiences. The act is supported through the use of props and costumes, such as the uniform required by rank or profession, the preferred fashions and styles of a social group, or the favoured brands of a hobby. The act must also be performed in an appropriate setting, in an appropriate manner. Manner encompasses aspects of the performance not done through props or costumes, such as behaviour and language. Overall, the success of the performance depends on both the skill of the actor and the willingness of the audience to accept the show, making tactful allowances where appropriate and to avoid being embarrassed, embarrassing others, or allowing others to be embarrassed.

While Jenkins emphasises identity as a process that is done by us *to* others and not a thing that is had or given, Goffman's framework permits a focus on the ways identity is done by us *for* others. As new technologies emerge, new ''social establishments'' develop. The flexibility of the dramaturgical framework has seen it applied to online social networks. These provide new roles to play, new audiences to perform for, new props and costumes to use. The internet offers not just one but infinite new stages upon which to perform. An early study of American internet users, conducted by questionnaire, found that different

technologies were preferred for different aspects (e.g. family, social, work) of a person's identity (Farnham and Churchill, 2011). People, when asked to rate the compatibility of their online identities, varied in the degree to which they felt these were integrated or faceted. This incompatibility resulted in increased self-monitoring of what was shared on different platforms.

Different platforms naturally offer different ways to perform identity and express the self: profiles, avatars, banners, status updates, friends, followers. To Goffman's theatre, Hogan (2010) adds an art gallery --- social media users not only perform the self in front of a live audience, they also stage exhibitions. These carefully curated uses of website functionality depend not only on the user's relationship with their imagined audience, but on the user's relationship with their own identity. Identity is not only faceted but these facets are mutable. People at transitional points of their life can use social media as ''social transition machinery'' (Haimson, 2018). This machinery allows people to experiment with, develop and reconstruct their identity. The transition into adulthood increasingly takes advantage of this machinery (Jordán-Conde *et al.*, 2014; Kapidzic and Herring, 2015; Wood *et al.*, 2016) as well as going to college (Morioka *et al.*, 2016) or gender transition (Haimson, 2018).

How people perform identity through emoji has been examined in two different ways. First, there are the oblique studies. By oblique, I mean that these studies gather data from a population, divide that sample based on user demographics and then compare differences in emoji usage per demographic group (Barbieri and Camacho-Collados, 2018; Coats, 2018). While this says something about differences in usage between, say, male and female or young and old, there is no guarantee that this usage is entirely an act of identity expression. However, the frameworks of Jenkins and Goffman each allow for identity to be something understood from both intentional and unintentional acts. Chen *et al.* (2018) used data from the Kika Keyboard app to analyse differences in emoji usage between

female and male users who had self-reported their gender in the app. Some emoji were more frequently used by one gender, leading the authors to claim ‘‘the existence of female emojis and male emojis’’ and show that machine learning models trained only on emoji could predict a user’s gender with an accuracy between 70% and 83%.

A more direct approach is found in Ge (2019), which utilises discourse analysis (rather than large scale data analysis) and looked at sequences of emoji rather than individual emoji. This work measured the extent to which users on a Chinese microblogging site express their ‘‘textual voice’’, as reflected in the stance and engagement of their online posts containing emoji sequences. Self-identity in this work is grounded in variety of different works rather than Goffman or Jenkins, but the premise is similar: the presentation of self-identity is mediated through language, with people simultaneously trying to assert their individuality and connect with others. Ge found that emoji sequences are more often used to express stance (such as their attitude, or self-mentioning) than engagement (such as mentioning others, asking questions). However, this work is overly focused on a a small set of social media influencers and has little to say on any specific or demographic-based components of identity.

TME have been studied not in relation to performance of identity but their semantic properties (Coats, 2018; Barbieri and Camacho-Collados, 2018). These studies used large volumes of Twitter data to examine the sentiment associated with the five tone modifiers and the most similar emoji as determined through cosine distances applied to a vector space model, finding that the same underlying emoji can have different distributional patterns when combined with a tone or gender modifier. As with Chen *et al.*, while this does not directly connect emoji to identity it is still indicative of variation in usage that is plausibly due to different emoji being used to perform different identities. A more identity-focused approach is found in Li *et al.* (2020), but with no restriction to TME. They looked at

emoji usage in Twitter bios --- user-provided text which is shown in the header of every user's profile page and can be considered an exhibition as per Hogan (2010), uniquely positioned to express identity. Li *et al.* (2020) find that the emoji in a user's bio differ from those in their tweets. Bios are more likely to begin with an emoji and more likely to consist of *only* emoji. The emoji in bios are less likely to be facial expressions: heart, flag, sport and music emoji are more common in bios. Furthermore, the followers of these users tend to use similar emoji in their own bios.

2.4 Production, Perception and Experimental Methods

When Jenkins says the identity process is rooted in language, this will be of no surprise to sociolinguists. Decades of work has explored the relationship between variation in how we communicate and the significance of this in terms of not only who we are but who others *think* we are. Lambert *et al.* (1960) studied this at the level of language, using the Matched Guise Test. Participants evaluated recorded stimuli that differed only in the language they were presented in --- the content and the speaker were identical. In this case, a passage of French prose was translated into fluent English. Recordings of both were made by four French/English bilingual Canadians, selected for their fluency in both languages. Participant were non-bilingual Canadians and spoke either French or English. They listened to the recordings (along with non-critical fillers), unaware that these eight recordings were actually matched pairs and the only difference being the language spoken. Participants filled in a survey rating their perception of the speakers in the recordings on fourteen traits: height, good looks, leadership, sense of humour, intelligence, religiousness, self-confidence, dependability, entertainingness, kindness, ambition, sociability, character, and

general likeability. Contrary to the reasonable expectation that participants would more favourably evaluate speakers of their native language, both English and French participants comparably rated English speakers higher for many traits. Furthermore, French participants evaluated French speakers *less* favourably than the English participants did.

Lambert *et al.* interpret the results as ''a reflection of the influence of community-wide stereotypes of English and French speaking Canadians'', though do not attempt to say what (if any) specific properties of language are associated with any given trait. Work along those lines began with William Labov, whose studies of speech communities in the United States laid the foundations of variationist sociolinguistics Labov (1972). Early studies (Labov, 1986) explored how differences in the phonetic properties of speech used by people related to some social characteristic, such as class, as opposed to the attitudinal traits used by Lambert *et al.*. Labov had noticed differences in the extent to which people in New York pronounced the consonant *r* after a vowel (e.g. *car*/*card*, *four*/*fourth*). This difference seemed to be dependent on social class. To test this, Labov visited department stores in New York (Labov, 1986). Three were selected as ranked by perceived prestige, based on which newspapers they advertised in and the price of their wares: Saks Fifth Avenue (highest prestige), Macy's and S. Klein (lowest prestige). The assumption was that this stratification in prestige would also extend to employees and that this would manifest itself in the pronunciation of *r*. In each store, Labov asked sales assistants where a particular item could be found. Upon being told ''Fourth floor.'', he asked for confirmation by saying ''Excuse me?'' to elicit a more careful, emphatic, version of the phrase. Indeed, realisation of *r* was stratified: the higher the prestige of a store, the more often the consonant was pronounced by sales assistants. The consonant *r* is an example of what Labov labels a *sociolinguistic variable*: a specific linguistic variation in speech which *indexes* a particular *social variable*. Such indexicality operates

on different levels of perception and Labov categorises sociolinguistic variables either as *indicators*, of which speakers have no awareness, *markers*, of which speakers have only implicit awareness, or *stereotypes*, of which speakers have explicit awareness. The indexicality of speech has been widely studied beyond social class --- listeners use phonetic cues to infer aspects of speaker identity such as sexuality (Smyth *et al.*, 2003) but also more nebulous concepts such as how intelligent (Campbell-Kibler, 2009), trustworthy (Kinzler *et al.*, 2011), credible (Lev-Ari and Keysar, 2010) or prestigious (Beinhoff, 2013) a person is. Prior work has shown how these perceptions of a speaker, based purely on how they sound, can result in loss of opportunity (Baugh, 1996; Hopper and Williams, 1973).

Labov's awareness of the connection between r and social class arose from his experience of interviewing different communities, but does not examine whether those communities were aware this specific aspect of their speech. Campbell-Kibler (2009) discusses this awareness under the topic of ''social meaning'' and the idea that sociolinguistics is concerned with the relation between linguistic and social behaviour: ''linguistic variation not only reflects social differences, but is also used by speakers to position themselves within the social world, and through such positioning, to build and rebuild that world''. If emoji can be sociolinguistic variables, as this thesis argues, then when Twitter users put specific emoji in their bios (Li *et al.*, 2020), it is arguably because they have explicit awareness of using those emoji and also what they index, in terms of their identity. The same users can also rely on other people (at least, those people the user wishes to identify with) having that same explicit awareness. That emoji can have social meaning is the central claim tested by this thesis.

The Matched Guise Test is commonly used to investigate the social meaning of linguistic variables because it allows for fine control of stimuli and the removal of confounding factors. Forced binary choice experiments of this form, with no possibility to express uncertainty, are widely used in experimental psychology

where they are more commonly known as two-alternative forced-choice tasks. The method is used to understand decision making processes, with a dependent variable such as accuracy or reaction time (Ratcliff and Rouder, 1998; Bogacz *et al.*, 2006). The use of a binary choice with no option to express neutrality or uncertainty makes it possible to detect small effects when results are aggregated over participants. This is especially important for detecting the effect of sociolinguistic variables which operate at a less conscious level than Labov's stereotypes. Lambert *et al.* (1960) manipulated speaker language, while holding speaker identity language constant. Campbell-Kibler (2009) digitally manipulated speech to differ only in whether words ending in *-ing* have that syllable articulated as *in* or *ing*, while holding constant both the content of the speech and all other phonetic factors such as speed, pitch, volume and stress. The characteristics of participants can also be controlled for, as is standard in experiments --- Lambert *et al.* distinguished participants by their native language, Campbell-Kibler by gender.

In Chapter 4 I use the Matched Guise Test to explore the social meaning of emoji and the identities people associate with TME. The test is administered entirely online, using written stimuli rather than audio. The use of written stimuli in this way is neither novel nor controversial.

Bradac *et al.* (1988) manipulated lexical diversity in written conversations, showing that readers were able to detect whether speaker A's style converged or diverged over time (whether to a large or a small extent, or from more to less complex) to match the style of speaker B. Particular combinations of change type, extent and complexity resulted in speaker B being seen as exhibiting solidarity, higher status and more competent.

Buchstaller (2005) manipulated the quotative use of *like*/*go*[2], to explore whether explicitly stated attitudes towards these different constructions, such as age, gender or class, correlate with judgements made in controlled experiments.

[2]"He was like, we should leave now" vs "So he goes, we should leave now"

While particpants overtly considered *like* to be used mainly by young female speakers, overt attitudes as measured through the Matched Guise Test found only a significant connection with age. The use of quotative *go*, thought to be stereotypically associated with lower class male speech, showed neither overt nor covert association with speaker characteristics.

Staum Casasanto (2010) examined *t/d* deletion, a form of linguistic variation which results in spoken phrases such as *fast car/band practice* sounding more like *fas car/ban practice*. Orthographically, the deleted part of the word can be represented with an apostrophe. As *t/d* deletion correlates highly with speakers of both African American Vernacular English and also non-vernacular Black Englishes, Staum Casasanto reasoned that *t/d* deletion may therefore be more associated with Black identities in general. This was manipulated in a Matched Guise Test. Participants did not fill in a questionnaire with questions on social variables. Instead they chose between two photos (showing a Black male and a White male) to indicate who they thought more likely to have said the sentence shown. The methodology of Staum Casasanto (2010) is of particular relevance to Chapter 4, where I will show that the social meaning of skin-toned emoji is shared between authors and readers.

2.5 Identity and behaviour

The identification process, when coupled with racism, sexism, ageism or any interpersonal ‘‘phobia’’, can result in negative outcomes for one party. Positive outcomes are also possible, if identity aligns with some personal or societal preference. These situations rely on the *entire* identification process, of which linguistic cues are but one component. However, that people consciously or otherwise connect variation to particular identities leads naturally to the question of what effect this has on behaviour and whether only particular types of cues

(e.g. linguistic, clothing) can trigger such behaviour. Prior work suggests this can indeed be the case.

Names can be a highly salient signal of ethnicity. Ge *et al.* (2016) studied the relationship between a person's name and their experiences on ride-sharing applications such as Lyft and Uber. Study passengers were assigned a name strongly associated with ethnicity (Black American/White American, e.g. Darnell/Brad), along with a matching profile photo. When requesting a ride through Lyft, drivers can see the passenger's profile photo while Uber drivers see no information until the journey is accepted. Ge *et al.* tested whether perceived passenger ethnicity affected waiting times. For both services (regardless of other factors such as gender and route characteristics) passengers with Black American names waited longer for their ride requests to be be accepted. For Uber, those passengers waited 30% longer for their rides to arrive after having been accepted by the driver.

The connection between identity signals and behaviour has been studied in experimental economics, using the Trust Game. In this game pairs of anonymous players are designated sender and responder. The sender has a fixed amount of money and can keep the whole sum or send a portion to the responder. Any money sent this way will be tripled. The responder may keep everything they are sent, or return a portion back to the sender. Pairs play one round consisting of the sender making their decision, then the responder making their response. Buchan *et al.* (2008) found an effect of gender on trust (as measured by the amounts sent/received), with female senders less likely overall to send money, but to send more to male responders than to female. Stanley *et al.* (2011) found a relation between the extent of a person's racial bias (as measured through the Implicit Attitudes Test (IAT) (Greenwald *et al.*, 2002)) and the extent to which they trust players of a different racial background. However, it should be noted that there is disagreement over the extent to which the detection of such biases, through the IAT, is correlated with specific behavioural outcomes (Blanton *et al.*, 2009)

as well as issues with replicability (Harris *et al.*, 2013).

More recently, Babin (2020) examined the effect of emoji when trust game players are allowed to communicate. Using emoji increased trust and reciprocation between players, but of particular relevance to this thesis is the finding that use of skin-toned emoji may result in less trust when those skin tones are darker. Babin did not set out to evaluate the impact of skin tone and so this result must be regarded as preliminary. However, it is suggestive of skin-toned emoji being able to induce bias to the extent that behavioural outcomes are affected and this therefore motivates the experiments presented in Chapter 5.

2.6 Terminology

Throughout this thesis I use terms to refer to groups of study participants based on their identity. I am keenly aware that the terms ‘‘Black’’ and ‘‘White’’ do not have a universally agreed-upon interpretation, since these are socially constructed categories (Ogbonnaya-Ogburu *et al.*, 2020; Delgado and Stefancic, 2017). Their precise meaning to different individuals, and the extent to which individuals identify with these labels, may vary even within a specific region such as London. Similarly, even within London, many different dialects of English are spoken, and linguistic factors can signify not only ethnic identity but also age, class, and other social variables. Indeed, it is this very possibility which motivates this thesis in the first place.

I use the terms ‘‘Black’’ and ‘‘White’’ to refer to the self-identified ethnicities of participants and the categories to which they assign imagined authors. Although there are popular dictionary-derived delineations between the terms ‘‘race’’ and ‘‘ethnicity’’ (e.g. race is physical, ethnicity is cultural (Blakemore, 2019)), these obscure the socially-constructed nature of both, as considered within the framework of Critical Race Theory (Ogbonnaya-Ogburu *et al.*, 2020). I use

the term ‘‘ethnicity’’ here for consistency with the self-identity question on the platform used to recruit study participants, which also referred to ethnicity.

Chapter 3

Production

This chapter consolidates work published in two papers: ''Self-Representation on Twitter Using Emoji Skin Color Modifiers'' (Robertson et al.*, 2018) and ''Emoji Skin Tone Modifiers: Analyzing Variation in Usage on Social Media'' (Robertson* et al.*, 2020).*

As noted (Section 2.2) the introduction of TME in 2015 was not uncontroversial and resulted in a range of claims regarding how these new emoji would be used and by whom. The major contributions of this chapter are an evidence-based analysis of these claims and the finding that TME are most often produced by users in service of self-representation. This finding raises question of whether these acts of self-representation by authors are understood as such by readers, which is the focus of Chapter 4.

This chapter offers a data-driven analysis of what I refer to as emoji *production*: the observed patterns of TME usage by users on social media. The research questions this chapter addresses are as follows:

- Compared to the default yellow, how common are tone-modifiable emoji?
- Is usage of the five available tones distributed equally, or are some more common than others?

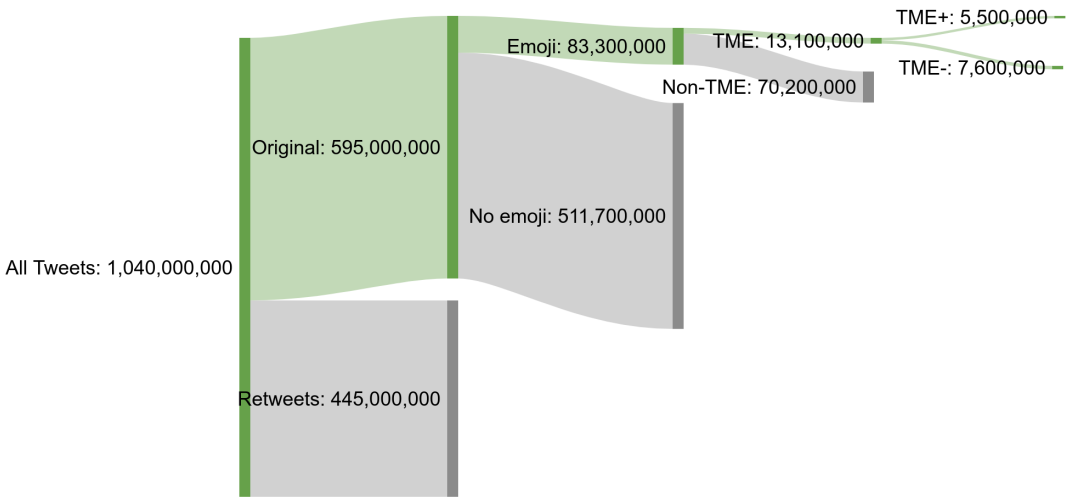


Figure 3.1: Breakdown of tweets in terms of their emoji content.

- Do tone distributions differ by geographic region?
- Do people use emoji tones which match their real-life skin tone?
- Where do people fall on a spectrum of never using emoji tones to always using emoji tones?
- How prevalent is "digital blackface", the practice of using darker emoji skin tones for abusive purposes?

3.1 Tweet Dataset and initial analysis

The Tweet Dataset comprises 1.04 billion tweets collected with the Twitter 1% Sample API between February 2017 and December 2017. Duplicate and retweeted tweets were removed, leaving 595 million unique original tweets. This data is used for an initial gross analysis of emoji skin tone modifiers.

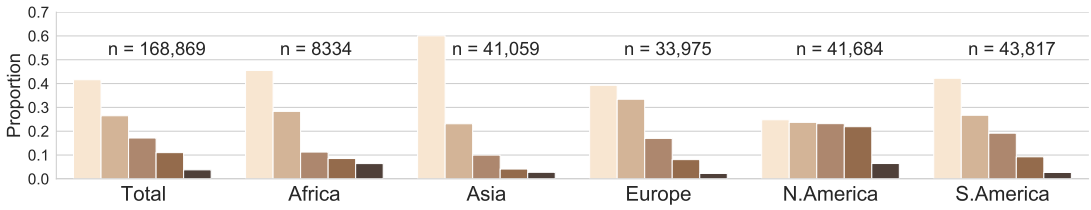


Figure 3.2: Proportion of TME+ skin tones, by geographic region.

3.1.1 How common are TME?

The extent to which different classes of emoji are observed in the Tweet Dataset is shown in Figure 3.1. 83.3 million tweets contain at least one emoji. Of those tweets, 13.1 million contain at least one TME and 5.5 million of those tweets with TME contain at least one TME+. Skin-toned emoji are therefore found in 0.92% of all original tweets in the Tweet Dataset. The approximate daily volume of tweets in 2017 based on the 1% sample was 350 million and, while 0.92% may seem miniscule, at the scale of Twitter this represents many millions of tweets per day that contain emoji to which a user has applied a skin tone modifier. Tweets containing TME- outnumber those with TME+, but the ratio of the two is not wildly imbalanced --- approximately 1.38 tweets with TME- for every tweet with TME+. This estimate is conservative with regard to TME- as it only takes into consideration tweets which contain either TME+ or TME-, with tweets containing both being counted towards TME+ only. Overall, this result stands in stark contrast to claims that TME+ would simply not be used (Zimmerman, 2015).

3.1.2 Distribution of TME, overall and by region

Of the 5.5 million observed TME+ in the Tweet Dataset, lighter tones dominate: 68% are either Tone 1 (T1) or T2. T1 is prevalent even in regions were people with such skin tones are a minority and one may naïvely expect emoji tone distributions to more closely reflect the region's population. It may instead reflect lower levels

of internet access in countries within those regions, with access[1] being almost universal in countries such as the US (89%) and Canada (95%) but more limited in others, such as Sierra Leone (17%) and Zambia (19%). Furthermore, access to the internet may be further restricted for some ethnic groups due to political oppression (Weidmann *et al.*, 2016), while poverty (Frankfurter *et al.*, 2020) is an especially common factor in lack of internet access.

However, the claim that white people do not use skin-toned emoji (McGill, 2016) was based on analysis of 18,000 tweets made in the US and so may only hold for US Twitter users, as acknowledged by McGill: ''It's worth noting the unpopularity of white emoji tentatively appears confined to the United States.'' who then goes on to claim, without substantiation, that ''[e]lsewhere in the world, including the Middle East, white emoji are more common.''

To explore this possible geographic variation, I used the location field of the 5.5 million TME+ tweets in the Tweet Dataset. This user-provided field is rarely populated (only 3.1% of these tweets) and subject to noise (users can enter anything in the field), but can be validated using the Google Maps API and grouped into broad geographic regions. Doing so reveals (Figure 3.2) that T1 is the most common *everywhere*, even in America.

What might account for the overwhelming popularity of the lightest emoji tone, in the face of multiple claims that it would not be used? One possibility is that the demographic make-up of Twitter plays a major factor. Since access to social media requires resources (an internet connection and a relatively modern device), it is likely that simply more white people have access to Twitter. This holds true beyond Twitter and extends to online communication in general. Based on frequency data I have access to (via the Unicode Consortium) which is based on operating system-level input data rather than any one specific online platform, if the five emoji tones were counted as individual stand-alone emoji then their

[1]Based on the World Bank's Telecommunication Report, `https://data.worldbank.org/indicator/IT.NET.USER.ZS`

	Johannesburg		London		New York City	
	#	%	#	%	#	%
Black	3,389,278	76.42	1,088,600	13.32	2,088,510	25.74
White	544,530	12.28	4,887,500	59.79	3,597,341	44.34
Mixed	247,276	5.58	405,300	4.96	325,901	4.02
Indian/Asian	216,198	4.88	1,511,600	18.49	1,038,388	12.8
Other	37,545	0.85	281,000	3.44	1,062,334	13.1
Total	4,434,827		8,174,000		8,112,474	

Table 3.1: Aggregated demographics for three locations, taken from SA, UK and USA census data. Johannesburg is measured in terms of the greater metropolitan area, rather than the city centre alone.

frequency ranks would mirror the overall trend of fig. 3.2. Indeed, Tone 1 would be a top 20 emoji while Tone 5 would be only in the top 150.

However, McGill's claim is that *white people* would not use white TME+, not that *nobody* would use these emoji. To answer questions of *who* is using TME+, additional data is needed.

3.2 Users Dataset and further analysis

The Users Dataset contains 28.3 million tweets from 19,683 Twitter users selected based on geographical region, for whom the real-life skin tone was determined through crowd-sourced annotation of their profile photo. It contains an additional 43.4 million tweets from 26,937 users with no profile photo.

From 3.4 million tweets made by 2.6 million unique users, collected via the Twitter 1% sample API on the 14th of March 2018, I randomly sampled 50,000 unique users: 10,000 each from Johannesburg, London and New York City, and

20,000 random users. The additional random users were required as these profiles very commonly had no profile photo.

User location was determined using a more selective process than that described in Section 3.2: users were restricted to those likely to have used Twitter's auto-complete suggestions when personalising their profile, e.g. showing standard use of capitalisation and punctuation.

Johannesburg : the criteria was simply that location was listed as ''Johannesburg, South Africa''. 24,125 matches.

New York City : user location was one of ''New York, NY'', ''Brooklyn, NY'', ''Manhattan, NY'', ''Bronx, NY'', ''Queens, NY'' or ''Staten Island, NY'. 111,746 matches.

London : after removing all instances of ''West'', ''North'', ''South'', ''East'' and ''Greater'' from user locations, a regular expression (below) identified users in London and its boroughs. 93,466 matches.

```
^London$|^City of London|^City of Westminster|          //
^Kensington and Chelsea|^Hammersmith and Fulham|        //
^Wandsworth|^Lambeth|^Southwark|^Tower Hamlets|         //
^Hackney|^Islington|^Camden|^Brent|^Ealing|            //
^Hounslow|^Richmond upon Thames|^Kingston upon Thames|  //
^Merton|^Sutton|^Croydon|^Bromley|^Lewisham|           //
^Greenwich|^Bexley|^Havering|^Barking and Dagenham|     //
^Redbridge|^Newham|^Waltham Forest|^Haringey|          //
^Enfield|^Barnet|^Harrow|^Hillingdon                    //
```

Random : 20,000 users were randomly selected from those not already in any of the other three groups. Initially, 10,000 random users were sampled but far

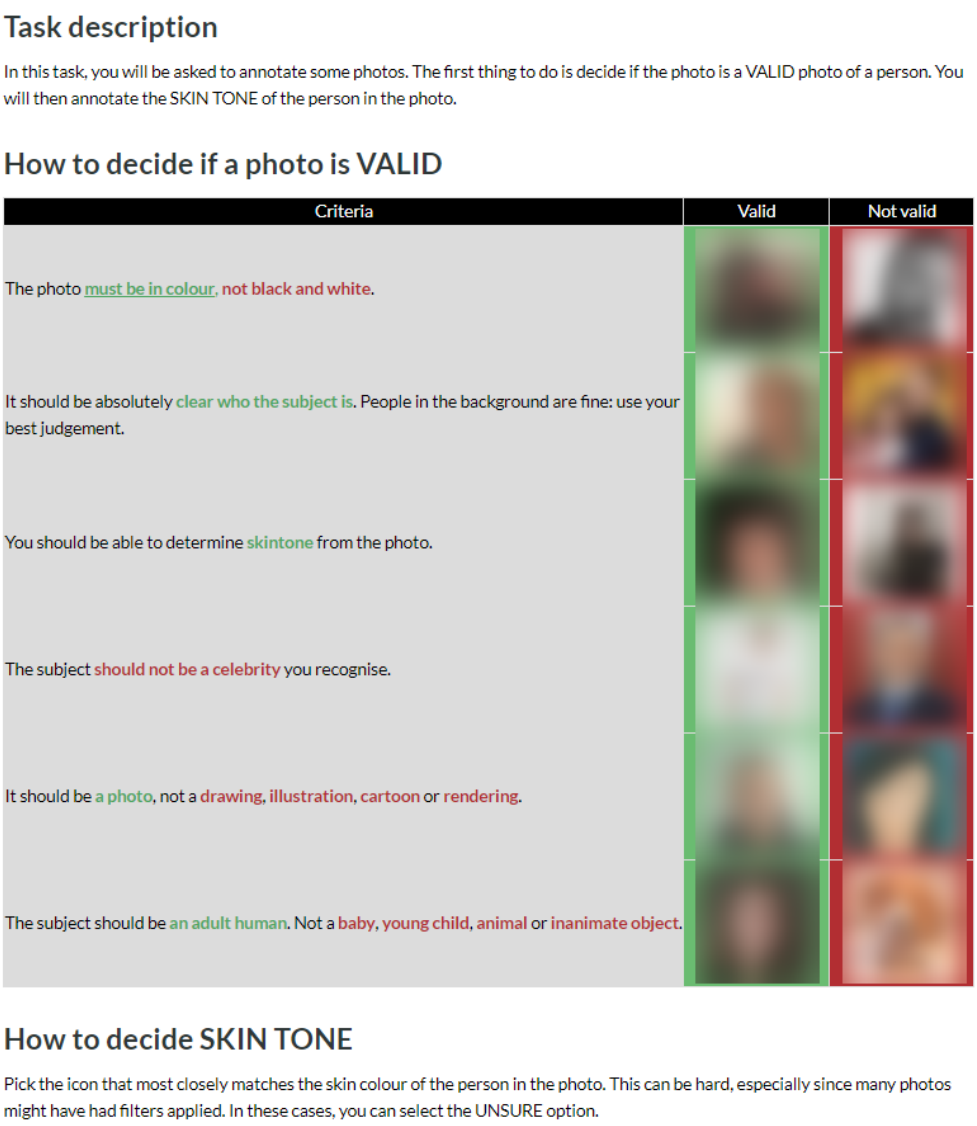
fewer of these users had profile photos and so the process was repeated to ensure similar numbers of users with photos in the four groups.

The three locations were chosen based on their demographic composition, each having different proportions of ethnic groups and therefore likely to have a range of skin tones. Aggregated census data supporting this is presented in Table 3.1. These locations were chosen not only for their varied demographic composition, but for the likelihood that many tweets would be available in English --- this allows manual inspection and analysis of the tweet content. To aid comparison between locations, sub-groups are collapsed into single categories where possible: e.g. Black African and Black Caribbean into Black; White British and White Irish into White. Johannesburg has a predominately Black population (Statistics South Africa, 2011), London predominantly White (Office for National Statistics, 2011). New York City is more balanced (United States Census Bureau, 2010). At multiple stages of the following process, users were removed from the dataset because their Twitter account had been deleted, banned or set to private. Accordingly, the initial 80,000 users were reduced to 46,620.

The profile photos were annotated using the Appen platform[2]. Annotators were shown a single photo at a time and instructed to first decide if the photo was valid and then state which skin tone best represented the user in the photo. A photo was to be considered valid if: it was in colour; it clearly showed the user as the subject; the user's skin tone could be determined; the subject was not a recognised celebrity; it was an actual photo and not an illustration; the subject was an adult human.

The guidelines given to annotators are shown in Figure 3.3 and were displayed along with every user photo. If the photo was not considered valid, the skin tone annotation options would not appear. Skin tone was shown using the ‘‘old man’’ emoji. This emoji was chosen because of the consistent hair colour, the

[2]`https://appen.com/`



Task description

In this task, you will be asked to annotate some photos. The first thing to do is decide if the photo is a VALID photo of a person. You will then annotate the SKIN TONE of the person in the photo.

How to decide if a photo is VALID

Criteria	Valid	Not valid
The photo must be in colour, not black and white.		
It should be absolutely clear who the subject is. People in the background are fine: use your best judgement.		
You should be able to determine skintone from the photo.		
The subject should not be a celebrity you recognise.		
It should be a photo, not a drawing, illustration, cartoon or rendering.		
The subject should be an adult human. Not a baby, young child, animal or inanimate object.		

How to decide SKIN TONE

Pick the icon that most closely matches the skin colour of the person in the photo. This can be hard, especially since many photos might have had filters applied. In these cases, you can select the UNSURE option.

Figure 3.3: Guidelines on photo validity shown to annotators. Example photos, which were manually selected from random Twitter users not used in this study, are blurred here to protect privacy.

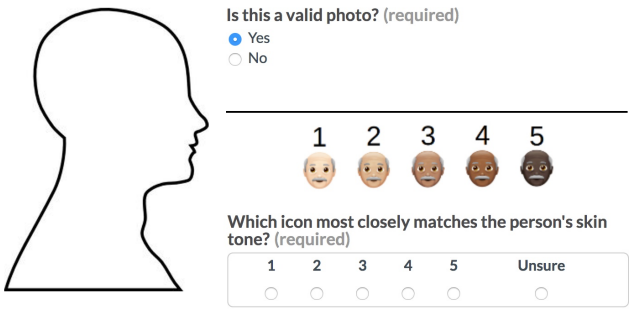


Figure 3.4: The actual annotation interface, with placeholder image.

	Users per skin tone group (with photo)						No photos
	1	2	3	4	5	**Total**	**Total**
Johannesburg	480	136	481	2,995	466	**4,558**	**5,065**
London	3,346	446	198	423	90	**4,603**	**5,093**
New York City	2,896	530	246	729	122	**4,523**	**9,091**
Random	4,085	1,063	287	479	85	**5,999**	**11,688**
Total	**10,807**	**2,475**	**1,212**	**4,626**	**763**	**19,683**	**26,937**

Table 3.2: Number of users per region, per skin tone group, with and without a Twitter profile photo.

lack of distracting visual features and the ease by which the emoji's skin tone could be observed. An option for being unsure about the skin tone was available. Test questions were included to filter out individual annotators who failed easy annotation examples which obviously did not meet the validity criteria. No test questions were included based on skin tone. Annotators who failed to achieve 100% accuracy on test items were not allowed to progress to the actual annotation task. Each photo was annotated by three different annotators.

The final dataset consists only of Twitter users where at least two annotators agreed the profile photo was valid and at least two annotators also agreed upon the skin tone. Fleiss's Kappa for inter-annotator agreement is commonly reported for this type of data but interpretation of the value is not universally agreed upon, other than higher is better. Separately, Kappa was 0.94 for photo validity and 0.56 for skin tone. A more illustrative measure is that 61.3% of user photos were considered valid by at least two annotators and, for those valid photos, 83.2% had at least two annotators agree on the skin tone. The mean standard deviation for skin tone annotations was 1.07: disagreement between annotators generally involved a difference of only one tone. Therefore, annotators disagreed on photo validity in many instances but when in agreement they also closely agreed on the

	Users with photos			Users without photos		
	Mean	Std	**Total**	Mean	Std	**Total**
Johannesburg	1,113	1,222	**5,074,264**	617	1,036	**4,932,450**
London	1,628	1,204	**7,493,712**	1,544	1,231	**7,930,530**
New York City	1,429	1,241	**6,456,190**	1,317	1,277	**6,740,760**
Random	1,551	1,408	**9,294,404**	1,984	1,802	**23,771,425**
Total			**28,318,570**			**43,375,165**

Table 3.3: Number of original tweets per location, for users with and without a Twitter profile photo.

skin tone of the user in the photo.

The outcome of the annotation process, after removing profiles which were no longer publicly available since beginning the construction of the dataset, is shown in Table 3.2 --- a total of 19,683 Twitter profiles annotated for user skin tone based on a valid profile photo, in addition to 26,937 profiles without any user photo.

For all users, either successfully annotated or with no profile photo used, their most recent tweets were collected using the Twitter API in April 2018. The API limits this to 3,200 tweets and so I repeated this process again in November 2018. Since skin tone modifiers became available for use in April 2015, tweets preceding this were discarded. Statistics on the number of tweets collected are shown in Table 3.3.

3.2.1 Do white people use white emoji?

Figure 3.5 shows the proportion of users have always, sometimes or never applied a tone modifier where possible, grouped by location and annotated skin tone. On the assumption that users with skin tone 1 (T1) are likely to be white, the Users

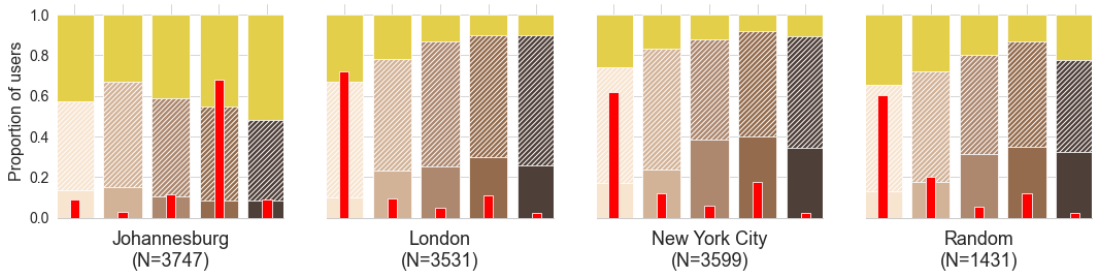


Figure 3.5: Proportion of users (grouped by location/skin tone) who produce TME as TME+: always (solid skin-toned), sometimes (hatched), never (yellow). Red bar denotes proportional population size for users in each subgroup.

Dataset provides evidence that white users, in all regions, *do* use TME+ and actually do so at similar levels to all other tone users --- on average across all regions, 68.9% of T1 users use TME+ sometimes or always. This is the same as all other tone users across all regions.

However, T1 users are least likely to *always* use TME+ at, on average, 13.4% while T2-T5 users range from 16.7% to 21.7%.

3.2.2 Do emoji tones reflect identity?

Figure 3.6: Proportion of most-used skin tone modifiers, by user skin tone and location.

Again, the claim of McGill (2016) was that white users in America do not use white-toned emoji. So far I have shown that the lightest tone is used in America (Figure 3.2) and that white users do use skin-toned emoji Figure 3.5. However, the real question raised by McGill is whether white people use white emoji. Indeed,

do people in general use TME+ that match their real-life skin tone?

The rows of Figure 3.6 represent users from each location, grouped by annotated skin tone, while the columns represent emoji tones. Each cell reports the proportion of users with real-life skin tone X whose most commonly used TME+ has tone Y: e.g. the bottom left corner shows users annotated as T5 who mostly used T1 emoji. Each row is normalised by the number of users per skin tone group for each location.

If people *only* use TME+ which match their real-life skin tone, then the left-right diagonal cells of Figure 3.6 should all have a value of 1.0. While this is not the case, the majority of users do use TME+ that match their own skin tone, particularly T1, T3 and T4 users. T2 users' usage is spread across emoji tones 1 to 3. T5 users stand out as mainly using the lighter T4, rather than T5.

Accordingly, in the majority of cases and even for white users, the choice of emoji skin tone reflects the identity of the user. I discuss the non-matching cases further in Section 3.3.

3.2.3 Are emoji tones used in abuse?

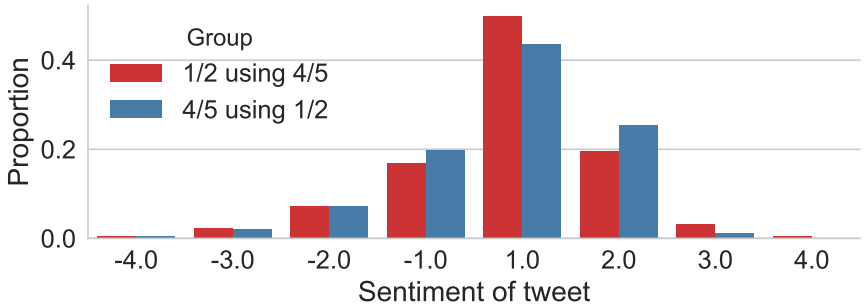


Figure 3.7: Distribution of non-neutral sentiment for English tweets containing lighter/darker TME+ but written by users having darker/lighter skin.

As feared by several media commentators (Dickey, 2017; Princewill, 2017), users may employ emoji skin tones which do *not* match their real-life identity,

as part of racist and abusive behaviour online. To explore this, I aggregated the Users Dataset into two groups based on user skin tone: tones 1/2 (lighter) and tones 4/5 (darker). For each group I selected only those English language tweets containing a TME+ of the ‘‘opposite’’ tone e.g. a user with skin tone 1 or 2 using a TME+ with tone 4 or 5 applied to it. These comprised very few tweets -- 0.14% of all tweets by T1/T2 users, 0.24% by T4/T5. The sentiment of these tweets was computing using Sentistrength (Thelwall *et al.*, 2012), a tool designed specifically for use with short social media texts, on the assumption that abusive posts are likely to contain negative language, such as slurs, which a lexicon-based approach such as Sentistrength would easily detect.

The majority of these tweets were neutral (almost 50% of the tweets of both groups). The distribution of the non-neutral sentiment of tweets for both groups are shown in Figure 3.7, where 4 is the most positive and -4 is the most negative 4. As shown, the distributions for both groups are almost the same, and positive tweets outnumber negative ones. Inspection of tweets with negative sentiment revealed tweets on generally negative topics (e.g. not enjoying going to the dentist), rather than anything specifically abusive.

Overall, there is no evidence to justify fears of widespread ‘‘digital blackface’’ or its black-against-white counterpart. It seems more likely that users wishing to be racially abusive online will resort to traditional abusive language to do so. Of course, such language may be augmented with non-TME emoji.

3.2.4 Profile photos and TME+ usage

Analysis so far has considered only those users who publicly display a profile photo considered valid (i.e. shows the user) by annotators. This may be a confounding factor in examining self-representation since these users are, by definition, engaging in self-representation from the outset.

Towards addressing this issue I compared users with and without profile

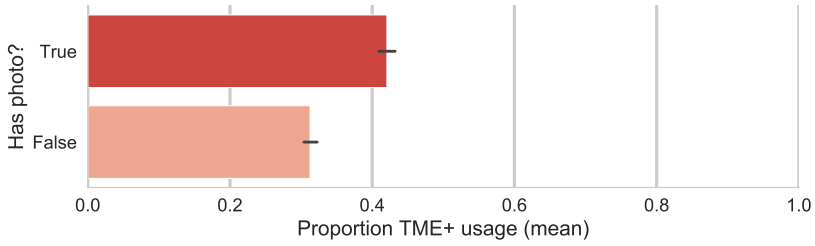


Figure 3.8: Average use of TME+ as a proportion of all TME by users with and without profile photos, with error bars showing bootstrapped 95% confidence intervals. Users without a profile photo use TME+ significantly less often.

photos, in terms of their TME+ usage, by dividing the random group based on the presence/absence of a valid photo on their profile. Since some user photos classified as invalid inthe Users Dataset (Section 3.2) may actually show the user, I reclassified those containing multiple people as 'with-photo'. This resulted in 6,188 (45.4%) users with a photo and 7,430 (54.6%) without. I then computed each user's proportion of TME+ out of all TME used, with the average for both groups shown in Figure 3.8.

The difference in mean TME+ usage between uses with a profile photo (M=0.42, SD=0.43) and without (M=0.31, SD=0.41) is significant (independent samples two-sided t-test: $t(13616)=11.8$; $p<0.001$). Users who engage in self-representation through a profile photo also use TME+ more often. More anonymous users, without self-identifying profile photos, do still use TME+ in their online messages, but less frequently.

However, a further analysis suggests that although the no-photo group use TME+ less often, when they do use TME+ they do so in similar ways to the with-photo group. Specifically, for each of the six TME configurations (tones 1-5 or no tone), I computed the proportion of users in each group who had ever used that TME configuration.[3]

[3]I did not use counts over all individual TME as this would be dominated by users who produced a lot of TME.

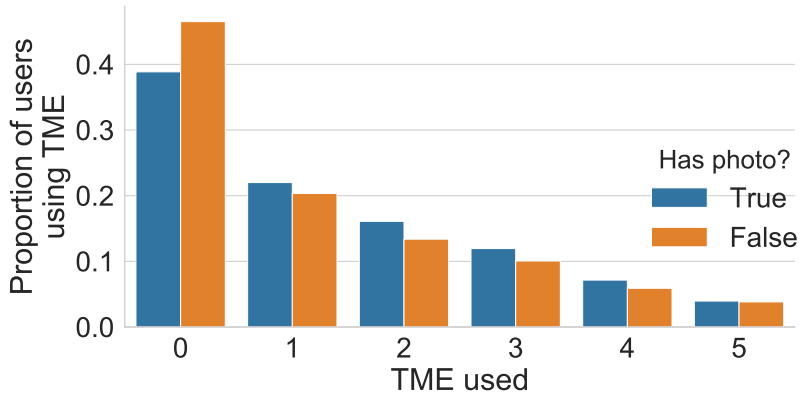


Figure 3.9: Proportion of random sample users, with/without a profile photo, who have used one of the six possible TME configurations: zero tone (yellow) or tones 1-5.

While the two groups differ in their overall use of TME- vs TME+, the distribution of TME+ usage across the two groups is strikingly similar (Figure 3.9). Although this does not directly confirm that the two groups are similar, either in terms of the distribution of their skin tones or their choice to use particular tones to represent themselves, taken together these two assumptions would be the simplest explanation for the observed match between the two groups' distribution over TME+ tones. The alternative would require between-group differences in these factors to precisely counterbalance each other in order to produce the observed pattern of results.

3.3 Variation in TME production

It is clear that users do not produce TME+ at every opportunity (see the cross-hatched bars of Figure 3.5), nor do they always choose an exactly-matching tone (see the non-diagonal values of Figure 3.6) when they do produce TME+. There are also group differences in usage: T5 users are slightly less likely to use TME+ and far more likely to use Tone 4 when they do, particularly in Johannesburg. These findings motivate moving beyond the claims investigated in Sections 3.1 and 3.2 and so in this section I more closely examine the observed variation in

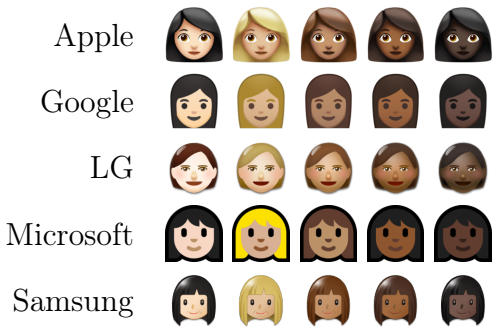


Figure 3.10: Examples of how TME+ are rendered on five vendors.

TME usage.

3.3.1 Variation in most commonly used tone

As shown in Figure 3.6, a user group's most-used emoji tone is not not always that which most closely matches perceptions of real-life skin tone, with T2 and T5 showing the greatest divergence. These differences in usage are likely due to the visual affordances and design of skin-toned emoji, specifically the hair and skin colours used. T2 TME+ always have blonde hair, as shown in Figure 3.10, so in regions where very dark hair is most common (e.g. some Asian countries (Leerunyakul and Suchonwanit, 2020)) T1 emoji may be preferred because these strike the best balance between hair and skin colour. Users might be avoiding this emoji tone because it does not match their hair, while other users may choose it *because* it matches their hair colour, even if it is *not* the closest match for their skin tone.

Furthermore, the visual differences between the five toned emoji on all platforms are based purely on changes in the skin/hair colour --- this leads to reduced visual detail in the eyes/nose/mouth due to lack of contrast between these features and the skin. This may make them less attractive to users. There is also the choice of colours made by vendors in their designs. The gradient from lightest to darkest emoji is not uniformly smooth, with the change from T4 to T5 being

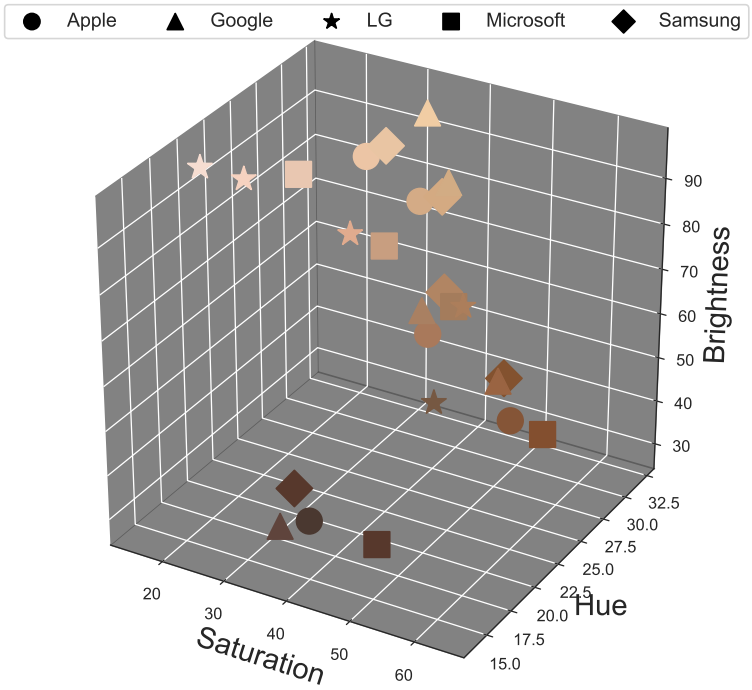


Figure 3.11: Hue, saturation and brightness values of predominate pixel colour in 🖐 TME+, for five vendors.

perceptibly greater than that of any other two adjacent tones. To quantify this observation, Figure 3.11 shows the distribution of each vendor's emoji tones, based on the hue, saturation and brightness (HSB) of the most common pixel colour in each hand emoji. The HSB model is an alternative to RGB, designed to reflect the way the human visual system produces colour sensation (Joblove and Greenberg, 1978). For most platforms, the darkest tone T5 is an outlier. The exception is tones used on LG phones prior to 2017[4].

The impact of how emoji skin tones are implemented is twofold. First, T5 users are less likely to use TME+ and those who do are more likely to use T4. Second, users may have to chose between representing their hair or their skin.

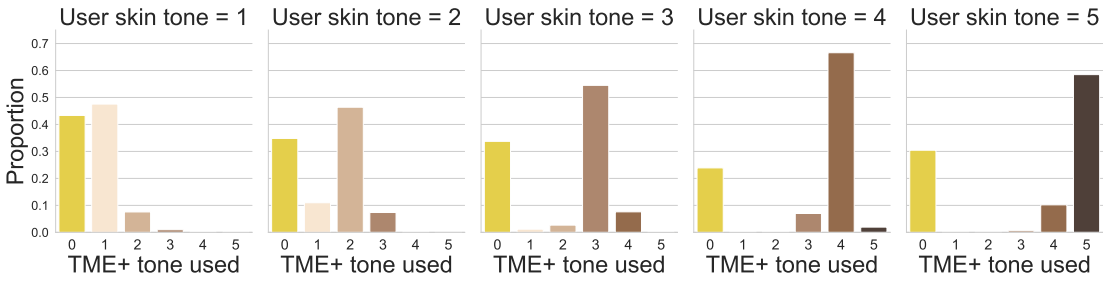


Figure 3.12: Evidence for tone modulation. Distributions of tones used in TME+, plus TME- shown in yellow, by users (aggregated by skin tone) whose most common TME+ tone matches their actual skin tone. Users produce a variety of tones, besides those which match their own identity.

3.3.2 Tone modulation

Analysis has so far focused on which tone is employed by users in the *majority* of their posts. This overlooks the fact that users actually produce a *variety* of tones. Figure 3.12 shows the distribution of TME+ produced by those users whose majority TME+ tone exactly matches their real-life skin tone. These are the users in the diagonal cells of Figure 3.6, aggregated across all datasets. As expected, the most common TME+ tone per group matches the skin tone of that group. Nevertheless, the distribution over all tones indicates that there is a degree of inconsistency in the tones users employ in individual TME+: evidence that users *modulate* their tone production. Furthermore, the second-highest proportion in each group is TME-, which is evidence that users may turn tones on and off completely. This is surprising, since platforms such as iOS and Android remember the last tone used for any given TME, while Twitter's web interface even allows users to globally set the tone for *all* TME.

An alternative explanation for the high proportion of TME- is that individual users started by never using TME+ and then at some point switched to using only TME+. Examining random users, however, reveals several different patterns

[4]It is possible that if I had used the LG tones for annotation, there would be a stronger match between users' perceived skin tone and their most commonly-used TME+ tone.

in their tone usage. Figure 3.13 shows variations in tone usage for a given emoji by five users, with each user being from one of the five different skin tone groups. These users, and potentially others, do in fact switch between TME+ and TME-, as well as between different tones, even when using the same base emoji.

This usage is characterised as follows. For a particular TME e, a user u generates a sequence of events E. Each event is a tuple consisting of e and an associated skin tone t_i:

$$E = [(e, t_1), (e, t_2), \ldots, (e, t_n)]$$

Tone modulation occurs when it is the case that $t_{i-1} \neq t_i$, for some fixed emoji e. The extent to which tone modulation occurs is quantified by examining all E generated by all users and counting instances where $t_{i-1} \neq t_i$. These instances can be divided into two groups: tone-tone (where a user changes tones between TME+), and on-off (where a user switches between a TME+ and TME-, or vice versa).

The proportion of tone modulations involving each TME, with and without tones, is shown in Figure 3.14. These data are aggregated over all four datasets, then divided according to user skin tone group. Rows indicate a TME's prior tone at t_{i-1}, columns indicate the tone applied when next used at t_i. For example, cell (0,5) shows the proportion of modulations from yellow to the darkest tone. The diagonal (instances of no change) has been removed in order to highlight modulation and is not included in the total number of modulation events used to normalize the proportions in each matrix (listed as ''total mod events''). The tone modulation percentage is the percent of all events that involve tone modulation (i.e., total mod events divided by all events). Each matrix cell shows the proportion of a particular tone modulation in terms of tone before change (rows) and after change (columns). All forms of modulation combined account for 3.89% of events in TME sequences. This makes it a relatively rare phenomenon.

Section 3.3.2 aggregates the statistics of each matrix shown in Figure 3.14,

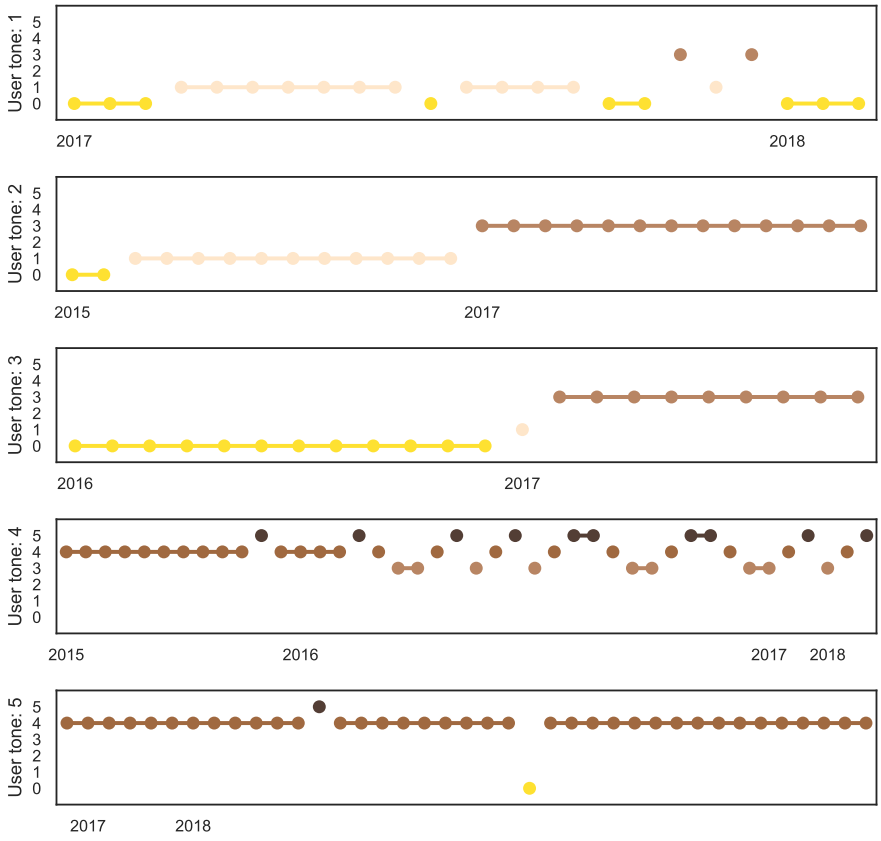


Figure 3.13: Five examples of tone modulation in users with different skin tones. Each row illustrates the tones applied to a specific emoji, each time that emoji was produced by the user. For example, the bottom row shows a user with skin tone 5, who generally applies tone 4 to 🙌, but for particular usage in 2018 applied tone 5 and for another turned the tone off.

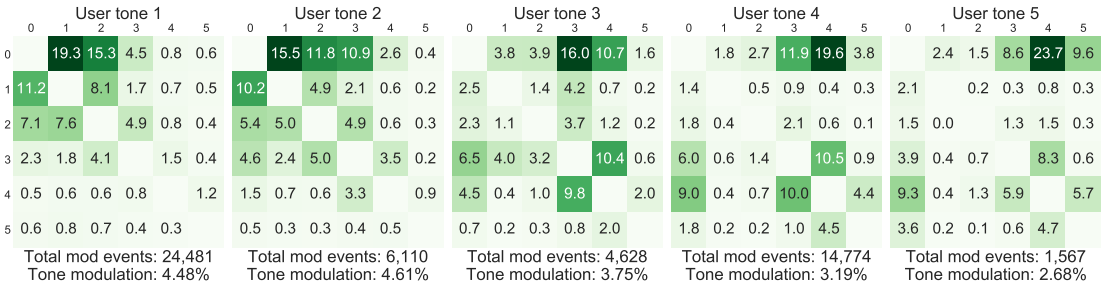


Figure 3.14: Proportion of tone modulations by user skin tone group.

User skin tone	Total events	Tone-Tone	On-Off	Off-On
1	24481	9299 (38%)	5292 (21.6%)	9890 (40.4%)
2	6110	2239 (36.6%)	1351 (22.1%)	2520 (41.2%)
3	4628	2198 (47.5%)	765 (16.5%)	1665 (36%)
4	14774	5959 (40.3%)	2951 (20%)	5864 (39.7%)
5	1567	529 (33.8%)	321 (20.5%)	717 (45.8%)

Table 3.4: Summary statistics for forms of tone modulation, based on Figure 3.14 data.

grouping tone modulations into tone-tone (where both t_{i-1} and t_i are TME+: all cells except those in row 0 or column 0), on-off (where t_{i-1} is TME+ and t_i is TME-: all cells in column 0), and off-on (where t_{i-1} is TME- and t_i is TME+: all cells in row 0).

From Figure 3.14, the single most common event for each user group is an off-on modulation from TME- to the TME+ which most closely matches their real-life skin tone, as seen in cell $(0, s)$ of Figure 3.14 where s is the user's actual skin tone. This may be due to users gradually enabling skin tones for individual emoji as they come into use. In general, tone modulation is clustered around the tone associated with the user's real-life skin tone --- tone-tone changes involving large differences are rare. Tone-tone modulations account for 30-40% of all modulations, with about the same proportion for off-on, and the remaining 20% or so for on-off (Section 3.3.2).

The extent of tone modulation suggests it is not merely accidental. Accidental changes in tone are certainly possible but, given the design of most common interfaces for inputting TME, somewhat improbable: on iOS, for example, producing 🖐🏾 after having last produced 🖐🏻 involves pressing on 🖐🏻 for a fraction of a second, then dragging to select 🖐🏾 and releasing. It seems unlikely that such an accidental

sequence of events can explain all of the observed modulations.

Another explanation considers that users can have multiple devices. They may therefore either have not selected a tone on one device (resulting in on-off events) or have selected a different tone on each device (resulting in tone-tone events), perhaps accidentally or due to the platform rendering differences seen earlier in Figure 3.10. By constructing E as a sequence of triples (e, t_i, p_i) where p_i represents the platform used for that event (provided by the ''source'' field as returned by the Twitter API), the number of tone modulations involving two different platforms (where $p_{i-1} \neq p_i$ as well as $t_{i-1} \neq t_i$) is approximately 16,000, around one third of all tone modulations. Even under the assumption that *all* cross-platform tone modulations are an artefact arising due to the use of multiple devices, and including all of users' one-off tone modulations as accidents, there remain a large number of users generating tone modulation events which are unexplained.

To characterise tone modulation events in more detail, I examine a subset of tweets and manually classify them based on overall observations, with the aim of identifying factors which precipitate tone modulation without making any prior assumptions as to what those factors might be.

Overall, tone modulation is an uncommon event, affecting 4% of TME observed in 28.3 million tweets. The majority of these cases involve tone-tone modulation, rather than the on-off variety. The major precipitating factor of tone modulation in all cases proves to be reference to other people, either directly by username or real name or indirectly through deixis. There is also evidence that social media users choose to leverage specific visual properties of certain emoji, particularly hair, even when other properties are not congruent with other aspects of their own appearance.

Where is the Unity hiding at???? Smh. 😔✊🏿✊🏾✊🏻😢

Don't feel it with conformity remain the exception ✊🏻✊🏾✊🏿

#BlackWomenMatter ✊🏾👸🏾👸🏾👸🏾

Figure 3.15: Examples of "tone rainbows" observed in tweets.

3.3.2.1 Tone-tone modulation

Instances of this form of tone modulation can be observed in Figure 3.13, where emoji vary between multiple possible tones. To investigate these events, I selected all tweets made by the random user group where a TME+ differed in tone by at least 2 since the last use of that TME, and the new tone did not match their annotated skin tone --- for example, a user annotated with skin tone 2 using 👌🏻 at one time point but 👌🏾 at the next instantiation of that TME+. This choice was made on the assumption that a difference of two tones is unlikely to be an accident or random variation between two similar tones that both roughly match the user's own skin tone. The total number of such tone modulations was 1,341, out of 108,584 tweets containing TME+. 48% were made by T1 users, 16% by T2, 26% by T3, 8% T4 and 1% T5.

Tone modulation can occur either *within* or *across* tweets. 15% of tone modulations occur within a single tweet. I exclude these tweets from analysis because they are typically difficult to interpret: without looking at the wider context of the tweets, it is not always easy to determine the referrents of each individual emoji, if there even is one. An exception to this is "tone rainbows" --- tweets that apply multiple tones to the same TME in a row. These generally appear to be used to explicitly represent diversity or convey solidarity. The fourth row of Figure 3.13 comes from a user who has produced such a rainbow for the ✊ emoji, and further examples are illustrated in Figure 3.15.

Considering only the examples of cross-tweet modulation left 1,136 tweets.

Category	Count	% of tone mods
Direct reference	159	47.0
Oblique reference	61	18.0
Self reference	46	13.6
Group reference	10	3.0
Iconic reference	8	2.4
Miscellaneous	10	3.0
Indeterminate	44	13.0
Total	**338**	**100**

Table 3.5: Seven categories of tone modulation distinguished in tweets in the random user group.

Filtering out non-English tweets left 362 tweets. Conservatively assuming that all instances of tone modulation are accidental when there is a platform difference between tweets, the final set for analysis contained 338 tweets from 336 unique users.

Five clear categories of tone modulation emerged, plus a sixth category containing miscellaneous usages. A seventh category contains usages which are inscrutable. The proportion of these categories, relative to those tone modulations examined and to the total of TME-containing English tweets from which they were drawn, are shown in Section 3.3.2.1. The categories are now described in detail. Illustrative examples provided from the data are shown in Section 3.3.2.1.

Direct reference: The most common kind of tone modulation involves direct reference to other people. Direct reference involves using the real name of person, including their Twitter username or relevant hashtag in the tweet, or responding to a photo of a person. In all cases, the tone of the TME+ is similar to that of the person or persons being referred to (confirmed for Twitter usernames by

	UT	Prior TME+ usage	Modulated TME+
Direct	1	@ST1_user 👌🏻	@ST5_user @ST4_user Yeah buddy why not!!! 👌🏿
	1	Wow I'm gunna be 25 weeks this week already!! 👶🏽	I have a new niece, $realname 👶🏽💞
Oblique	1	✋🏻😊	I hope you get a paper cut ✋🏿🙄
	4	People Always Onna Outside Looking In Like STAY OUT MY MIX 🤘🏾	I can't stop a nigga from doing him & I won't try to either ✌🏽
Self	1	Go on early runs 🏃🏻‍♂️ w/ me so I know it's real	@ST2_user Me on pay day 🏦🏃🏻‍♂️
	3	if your ex still popping up in your notifications or life y'all got unfinished business so stay tf away from me 🤷🏾‍♀️	How somebody feel about me, ain't my business 🤷🏾‍♀️ that shit personal
Group	5	I'm done grabbing leaves cuz 🤦🏿‍♀️	White girls 🤦🏼‍♀️ .. #ILoveTheSistas
	4	Just got home. I'm so exhausted 😩😩😩 But very happy that my dad had a safe trip home 🙏🏾	If someone can just get me white privilege for Christmas 🙏🏽
Iconic	1	Chopped off 14 inches of my hair a year ago today & it's all back now 💁🏾‍♂️	Should I die my hair dark again or go lighter? 💁🏽‍♂️🎉
	1	@ST1_user yep because their fake ass bitches 👸🏽	Y'all Tuesday is the day 👀💁🏾‍♂️👯

Table 3.6: (UT=User Tone) Examples of types of reference made using tone modulated TME+. The modulated TME column shows tweet containing a TME+ which is at least two removed from the user's own skin tone. Real names and usernames have been replaced with placeholders which reflect the skin tone of the referent, where possible.

looking at the profile photo). The existence of this category is perhaps the most predictable---since TME+ are used for self-representation, the possibility that they can be used for representing others is a reasonable extension. What was perhaps less predictable was the differing extents of these two uses: it is extremely uncommon to use toned emoji to refer to other people with skin tones different from that of the user.

Oblique reference: These references are more vague, occurring in tweets which are not in response to other tweets. They contain no real names, usernames or hashtags which could refer to a person. Instead, they characteristically use deictic expressions such as ‘‘he’’ and ‘‘they’’. In some cases these expressions are themselves actually emoji, which is of interest as it suggests another possible source of evidence in support of claims regarding the representative power of emoji. This usage may be related to ‘‘vaguebooking’’, a practice whereby users post deliberately vague or ambiguous messages to social media, and which has been characterized by two seemingly diametrically opposed explanations: users post either to preserve privacy (Child and Starcher, 2016) or as a means of gaining attention (Berryman *et al.*, 2017).

Self reference: The referent in these tone modulations appears to be the user, though the TME+ skin tone is at least two removed from the user's tone, i.e. the user is annotated as T1 but the emoji used is T3. Use of personal pronouns is common and the emoji are often facial, especially 🤷‍♀️. Unlike direct and oblique reference, there are generally no usernames, real names, hashtags or other indicators of reference to others. Self-reference is further supported by the fact that these emoji can be gendered (in a similar way to how tones are applied) and in all cases emoji gender matches user gender, as determined by user profile photo. It is possible that these usages are input errors, given their rarity, but their co-occurrence with personal pronouns and matching emoji genders suggest this is unlikely to be the case in every instance.

Group reference: These TME+ appear to refer to a group or class of people. The target is non-specific, not targeting any particular person. In our observations, these tone modulations all refer to ‘‘whiteness’’ in some way --- white women, white privilege, geographic regions with a lighter-skinned population. The choice of TME+ tone is clearly deliberate and chosen with a specific purpose in mind.

Iconic reference: The visual affordances of some emoji encourage users to make reference to themselves via particular attributes of these emoji. In particular, these are based on differences in hair colour (as discussed in Section 3.3.1) and all involve the user's planned or recent changes to their hair colour. The only emoji used in this context are 💇‍♀️ and 💆‍♀️. The ‘‘hair cut’’ emoji is predictable, while the ‘‘face massage’’ emoji is likely used here because, on many platforms, it looks like someone having their hair washed by someone else.

Miscellaneous: Most instances of tweets in this category are song lyrics. In some, but not all, the skin tone more closely matches that of the song's artist. It may be that users have copied and pasted these lyrics, including the emoji, from elsewhere on the internet. A single instance of sexual use was observed, using T5 to refer to stereotypes.

@user 🏿

Wicked wicked wicked 🤘🏻

I'd rather fuck wit myself then waste my time 👌🏽💯

@user happy bday my boy 🤘🏽 much love bro 💯

Figure 3.16: Examples of more inscrutable tone modulation

Indeterminate: The remaining instances of tone modulation are somewhat inscrutable, not clearly fitting into the above categories. Examples are shown in Figure 3.16. In some, reference to another person is made but the tone used in the TME+ matches neither the user nor the referent. These may be input errors, as discussed earlier in Section 3.3.2. Or, in the case where the message is targeted at another user by including their username, there may be some private meaning to the use of these particular emoji. Since I examined only the preceding tweet rather than any full thread of tweets, it is also possible that some tone modulations in this group could have been classified more accurately into another group if more context were considered.

3.3.2.2 On-off modulation

Instances of this form of tone modulation can be observed in rows 1 and 5 in Figure 3.13, where emoji are realized as TME-, despite having been produced with a tone both immediately before and after. This is the least common form of tone modulation. This is shown by the first column of each matrix in Figure 3.14 and the summary statistics of Section 3.3.2.

To analyse this behaviour, I selected all English tweets made by the random user group containing a TME- where the author of the tweet had previously and subsequently used TME+ for that particular TME. I considered only those instances where the TME+ matched the user's skin tone, in order to select only

those instances of on-off modulation which appear to be a conscious decision to turn off tone. Instances where all three tweets were not published via the same platform (e.g. all from the Twitter app, or from the same phone vendor) were removed --- approximately 50%.

The total number of on-off tone modulations meeting these requirements was 75, out of 108,584 tweets containing TME+. 48% were made by T1 users, 22% by T2, 17% by T3, 7% T4 and 6% T5. Of the 75 tweets, only 60 were amenable to interpretation --- the remainder appeared to be from spam accounts with inconsistent TME+ usage.

Of these 60 cases, the majority involve reference to other people. In 18 tweets there is a direct reference by username where the author of the tweet and the person to whom they refer do not have a similar skin tone, based on observation of their profile photos. In 10 tweets, both parties did have a similar skin tone. In 3 tweets which included reference to multiple people, those mentioned had a variety of skin tones. Four tweets made generic reference to people at large and one made oblique reference to a person. There were 6 instances of reference to a Twitter account representing non-humans, such as businesses. In the remaining 14 tweets, which contain no usernames, the reference appears to be solely to the author. This may be due to using multiple devices of the same type, device reset or input error.

3.4 Reference vs representation

From the data observed, TME+ fit well within both Goffman's and Jenkins' conceptions of identity and self-presentation (Section 2.3). The overwhelmingly self-representational use, as evidenced by users selecting TME+ which closely match their real-life skin tone, makes TME a powerful yet flexible prop for expressing particular facets of identity online. TME are flexible not only for their

visual affordances (both skin and hair) but for the ease of which they may be selectively adjusted or even hidden. Tone modulation is the manifestation of that flexibility. This behaviour serves to aid others in what Jenkins calls the identification process, adding a cue for others to pick up on.

The existence of tone modulation suggests that there is a real question as to *when* users choose to send such a cue through TME+ usage. One factor may be the environment of the user and the salience of their skin tone to their identity within that environment. As noted in Section 3.2.1, users in Johannesburg showed the opposite trend to those in London and New York City --- Johannesburg users with lighter skins were more likely to use a matching TME+ than users with darker skin, and users with darker skin were especially likely to use only the default yellow TME-. From Figure 3.5, there appears to be a relationship between population size and TME+ usage: the population with the most common real-life skin tone produce TME+ less often than the population with the least common tone. One possible explanation is that users in a minority use TME+ to highlight, project or maintain this aspect of their identity. Language is commonly used to this end (see Gu and Patkin (2013) for discussion of language practices and attitudes in ethnic minority students in Hong Kong) and the internet provides spaces where users can develop, experiment with and express identities --- see works cited in Section 2.3. The role of emoji in these particular contexts remains an open question but the data here, in particular the variation observed based on geographical context, highlights the potential in exploring emoji in the ways that has generally been reserved for language or internet-specific modalities such as photos.

The relatively small number of cases of tone modulation in reference to others, and the fact that many cases of turning off tones entirely appear to be in reference to users with a different skin tone to the tweeter, suggests that some people may be unwilling to use TME+ to refer to others when there is not a common

racial/ethnic background. This could be due to the ‘‘networked public’’ (boyd, 2007) nature of Twitter. If TME+ are, in Goffman's terminology (Goffman, 1956), an important prop for self-expression then to be observed using the wrong prop could be viewed as inauthentic. This would account for the very low levels of tone modulation observed in the data, especially when considering the replicability issue of social networks like Twitter --- users may fear any past examples of such inauthenticity coming back to haunt them in the future. It could even be considered as misappropriation of another group's props and seen as an offensive act, similar to how particular words are restricted to in-group usage.

Misappropriated usage of toned emoji, if indeed offensive, would not show up as negative under the sentiment-based approach used in Section 3.2.3, which was purely based on lexical content. Therefore the extent to which negative racially-motivated communication involves TME+ could currently be under-detected. However, the methodology presented here for detecting tone modulation could be adapted to this purpose by relaxing the constraints I placed upon it in order to keep the qualitative analysis tractable.

3.5 Limitations

I note some limitations with the above methodology and analysis. These limitations, necessitated by having to reduce the volume of data for manual inspection, could affect the ability to accurately determine the full scope of tone modulation. In my analysis of tone-tone modulation, I only considered instances of maximal change in tone away from the user's own skin tone. This resulted in a very small sample and prevents being able to comment on cases where, for example, T2 users use T1 or T3 in their TME+. The qualitative results are therefore based on the most extreme examples of tone modulation. As a result, I do not distinguish non-self-reference in cases where the skin tones of the author and the referent are

the same. Therefore, any estimates of the extent of self-referential TME+ can be considered as an upper bound only. Future work should certainly attempt to determine the extent to which users of a given skin tone use TME+ of that same tone to refer to other people. While the results suggest that people avoid referencing another using TME+ tailored to the appearance of the other, it may well be the case that people are more willing to use non-self-referential TME+ when communicating with users with whom they share the same skin tone.

Another practical limitation is restricting the qualitative analysis to English tweets. Combined with only looking at the random sample, there may be cultural differences in tone modulation which have been obscured. This concern is somewhat alleviated by the results of Section 3.3.2, which showed there are at least no major differences in tone modulation between users based on their skin tone. Expanding the methodology to locations where skin tone carries especially strong cultural weight, such as South Asia (Baynes, 1997), could show the extent to which these cultural views are manifested through emoji usage in an environment where some tones may be considered undesirable.

In addition, many instances of tone modulation were not readily classifiable as references to other people. These appear to be self-reference but using a skin tone modifier that is unexpected, given the user's actual skin tone. Since I attempted to filter out accidental use by avoiding multi-platform inputs, some of these cases may involve additional precipitating factors which were not detectable given the method for manually inspecting tweets, which considered only the preceding and following tweets containing the target TME. These suggest some possible avenues for future work. A more detailed consideration of context might help, but ultimately determining the exact properties of the tone modulation phenomenon will likely require moving toward a more user-focused methodology. Targeted questionnaires and interviews may answer questions such as why a user chose a particular skin tone in a particular tweet, or why they chose to turn skin tones

off for one message in particular. This avenue of investigation is well-motivated by the results presented here, which show that tone modulation is a real aspect of TME usage.

Furthermore, the content analysis method used for this section deviates from best practices and its findings must considered with this in mind. There are many ways to undertake content analysis. One common approach is to employ multiple annotators who individually code a subset of examples. From this, a code book is constructed --- annotators discuss the identified types of data observed, reducing these to an agreed set which meaningfully captures differences in the data. This code book is then referred to when annotating the rest of the data, which is done by multiple annotators. Examples where annotators do not agree are reexamined and the degree of inter-annotator agreement is reported, through some measure such as a kappa score or correlation coefficient. Instead, I manually went through the data and labeled each example myself. I repeated this process multiple times, until I had the set of labels used in this chapter. As such, the result is overly personal and impressionistic and other readers (especially those more experienced in content analysis) may disagree with both the labels I produced and the examples to which I attached them. In any future work, I would certainly go to great lengths to avoid this.

Whether through big data methodology or more targeted interviews, a consideration of audience may be informative. Although Twitter suffers from context collapse (boyd, 2007), users can nevertheless control the size and nature of their intended audience to some degree by including usernames or hashtags in the tweet. Previous work has shown that Twitter users modulate their use of non-standard lexical items (another way of representing identity) based on audience size and type (Pavalanathan and Eisenstein, 2015; Shoemark *et al.*, 2017b,a). In general, a more restricted audience is likely to be more similar to the user who targets it, due to the homophilic tendencies of social networks like Twitter (Kwak *et al.*,

2010; Al Zamal *et al.*, 2012). This setting may encourage referring to other people by their appearance, especially if those others are of a similar ethnic background to that of the referrer.

Finally, note that the data examined here is from a single social media platform and is public in nature. Therefore, the findings here cannot speak to how users behave in private communication, such as in WhatsApp or Messenger groups or in private messages. In Goffman's terminology, this constitutes a different ''stage'' and therefore may elicit a different performance from people. I can only speculate as to how different these performances may be, if they are indeed distinct from public presentations. Future work could look at WhatsApp groups (Garimella and Tyson, 2018) which, while public, are arguably less so than Twitter. However, collecting data in this way raises ethical concerns, due to the blurring of public and private communications.

3.6 Chapter summary

The preceding analysis of Twitter posts, and the users behind them, resulted in the following findings:

1 TME+ were used in approximately 42% of Tweets where applying a tone was possible;

2 All five tones are used all over the world, with T1 being the most common;

3 Users of all skin tones use TME+;

4 Users who use TME+ choose a tone which closely matches their real-life skin tone;

5 Users with the darkest skin tone use a lighter emoji, due to poor visual design in T5 emoji;

6 The colour properties of the five tones are skewed towards the lighter range;

7 TME+ are not used alongside racially abusive language;

8 Use of TME+ and profile photos are linked, but users without photos still use TME+ in similar patterns to those who do;

9 Users selectively apply or remove emoji tones in specific situations;

10 Use of very different tones is uncommon and generally indicates referencing others.

TME, then, are easily characterised as a tool used to express one's personal identity. The scarcity of users producing TME+ with non-matching tones, the limited range of tones used by users, the lack of abusive usage, their popularity with users who also express identity through profile photos --- this usage follows from the self-representational nature of skin-toned emoji and places them neatly within Goffman's representational framework.

However, usage is necessary but not sufficient for a social understanding of these emoji: identity is not merely something we do with the tools at our disposal, but something that is in turn done to us by others who see those tools in use (Jenkins, 2014). This raises the question of whether the TME users produce have any effect on the identification process taking place in the minds of those who observe those skin-toned emoji online. This is the focus of the following chapter.

Chapter 4

Perception

This chapter is based on work published in the paper ''Black or White but never neutral: How readers perceive identity from yellow or skin-toned emoji'' (Robertson et al.*, 2021a).*

Chapter 3 provided evidence for the self-representational nature of skin-toned emoji. Since identity is most meaningful as a relationship with others, it is important to understand not only how people *produce* language and other artefacts to construct identity, but also how others *perceive* and *interpret* those acts of production.

Yet there has been little work in this area in terms of online communication. This chapter begins to fill that gap, inspired by work on spoken communication from the field of sociolinguistics as outlined in Section 2.4. That work shows that linguistic cues (ranging from detailed acoustic/phonetic characteristics to lexical and syntactic usage) can carry social meaning. Listeners use these cues to infer aspects of a speaker's identity, such as their regional background, social class or sexual orientation. Listeners also use linguistic cues to judge how intelligent, credible, prestigious or trustworthy a speaker seems. These cues alone are salient enough to provoke reactions to the speaker's perceived identity, such that the outcome can be negative and even harmful. Prior work has shown how this can

result in loss of opportunity for speakers whose identities are dispreferred by the listeners.

Focusing on the possibility that skin-toned emoji express an aspect of a person's ethnic identity, I hypothesise that individuals use emoji both to express their own identity and to understand the identity of others, as they do with spoken language. Furthermore, as has been demonstrated for spoken language, I hypothesise that this understanding can affect how readers react to content containing identity-related emoji --- for example, how likely they are to believe it or to pass it on. These potential behavioural consequences have clear practical implications in areas such as marketing, politics, and the spread of disinformation. In this chapter, I lay the necessary foundation for a follow-up study on behaviour (presented in Chapter 5) and ask:

- Do readers infer aspects of an author's identity (specifically, ethnicity) via their emoji usage (specifically, use or non-use of emoji skin-tone modifiers)?
- If so, do these inferences depend on the reader's own identity, and how does the emoji signal interact with other (linguistic) cues to identity?

Both the nature of social categories (here, ethnicity) and the potential indicators of those categories (here, emoji skin-tone modifiers and social media text) depend strongly on cultural context. Since this is the first work of its kind, I do not attempt a fully generalised study. Instead I limit the participant groups to native English-speakers living in London, to ensure a relatively common cultural and linguistic context, and focus my experiments on just two identity categories, Black and White --- both for participants' self-identified ethnicity, and the categories they were asked to infer.

This study design neglects participants who identify as neither Black or White, nor is it directly informative as to how emoji might be perceived to identify such individuals. However, for this initial study of emoji perception, I use the two

categories that I feel would be (to these participants) most typically associated with particular skin tones --- Black with the darker end of the Fitzpatrick scale, and White with the lighter end. My hypothesis (borne out by my experiments) is that the skin tones of emoji, like their real-life counterparts, are associated with particular identities, and are used by readers to infer the author's identity.

I am keenly aware that the terms ''Black'' and ''White'' do not have a universally agreed-upon interpretation, since these are socially constructed categories (Ogbonnaya-Ogburu *et al.*, 2020; Delgado and Stefancic, 2017). Their precise meaning to different individuals, and the extent to which individuals identify with these labels, may vary even within a specific region such as London. Similarly, even within London, many different dialects of English are spoken, and linguistic factors can signify not only ethnic identity but also age, class, and other social variables. Indeed, it is this very possibility which motivates this chapter in the first place. Nevertheless, I assume that participants will have some degree of shared understanding both of the categories Black and White, and of the linguistic features that tend to associate with those categories, at any level of awareness as per Labov's typology. Although there may be some differences between participants, the grouped design abstracts away from these differences. Investigating individual differences in the effects I report here is an important avenue for future work, but beyond the scope of my thesis.

Based on the above, I use the terms ''Black'' and ''White'' to refer to the self-identified ethnicities of participants and the categories to which they assign imagined authors. Although there are popular dictionary-derived delineations between the terms ''race'' and ''ethnicity'' (e.g. race is physical, ethnicity is cultural (Blakemore, 2019)), these obscure the socially-constructed nature of both, as considered within the framework of Critical Race Theory (Ogbonnaya-Ogburu *et al.*, 2020). I use the term ethnicity here for consistency with the self-identity question on the participant recruitment platform, which also referred to ethnicity.

I follow Blodgett *et al.* (2016) in referring to particular signals as ''demographically-aligned'', using the terms ''Black-aligned'' and ''White-aligned'' to refer to texts and emoji that are associated with these two groups. The alignment of texts is determined through a norming study (Section 4.2), and the alignment of emoji is based on the Fitzpatrick Scale labels assigned by the Unicode Consortium, with the two darkest tones considered ''Black-aligned'' and the two lightest ones ''White-aligned''. I do not use the middle emoji tone.

4.1 Hypotheses and research questions

This study tests three main hypotheses and several additional research questions inspired by the findings of Chapter 3 and the linguistic works discussed in Chapter 2. As noted in above, the hypotheses and research questions (as well as details of the experiment and analysis) were pre-registered with the Open Science Foundation.[1] The main hypotheses deal with how skin-toned or yellow emoji are interpreted by readers, when the emoji occurs with a *non-aligned* text: i.e., a text that is neither Black-aligned or White-aligned according to the norming study.

Hypothesis 1 Readers of a non-aligned text containing a Black-aligned or White-aligned emoji will more likely attribute authorship to a Black or White author.

This hypothesis tests whether readers use emoji as a salient cue to author identity. It is motivated by long-standing results from sociolinguistic studies showing that different social variables are indexed by specific linguistic variables (Labov, 1972) and that people are aware of these associations both consciously and subconsciously (D'Onofrio, 2018).

Hypothesis 2 Black readers will be more likely than White readers to attribute authorship to someone whose skin tone is similar to that of the

[1] Available online at https://osf.io/tn8mg/

emoji, given a non-aligned text containing a Black-aligned or White-aligned emoji.

The second hypothesis is motivated by the finding that people with darker skin tones produce more skin-toned emoji in their social media posts (Robertson *et al.*, 2018). I predict that this greater experience of expressing identity through emoji will result in Black participants in the experiment being more attuned to perceiving such emoji use, by others, as an explicit act of self-representation.

Hypothesis 3 Black readers will be more likely to attribute the default yellow emoji to a White than a Black identity, given a non-aligned text.

Claims that the yellow emoji is neutral and represents everyone run counter to evidence from Section 3.2.1, which shows that Black users in London are generally less likely to use the yellow emoji than White users. When presented with a text containing a yellow emoji, I predict Black users will more often judge the author to be White than Black. White users, who use the yellow emoji more often, should therefore see the yellow emoji as representing neither a Black nor a White identity.

In addition to these hypotheses are more general research questions. These compare the relative impacts of the identity signal encoded in text and emoji. Compared to a baseline of a Black-aligned or White-aligned text with no emoji, does adding an emoji result in a ''boosting effect'', wherein readers more often perceive the author as having an identity matching the combined text/emoji signal? Or is there no effect, possibly because one signal is already overwhelmingly salient? How do readers respond to conflict between the text and emoji signals, such as a Black-aligned text with a White-aligned emoji? Do they resolve the conflict by choosing one signal over the other, or do the signals moderate each other? Some differences between groups might be expected, since the two groups may not

be equally familiar with the linguistic characteristics of Black-aligned or White-aligned text, and Black readers may be more familiar with self-representational usage of emoji because of their popularity among Black Twitter users (Robertson *et al.*, 2018). Therefore, Black users may be more willing to ignore the text signal in favour of the emoji signal.

Research Question 1 What effect is observed when the identity signals encoded in text and emoji are congruent?

Research Question 2 What effect is observed when the identity signals encoded in text and emoji are incongruent, creating a conflict between the two?

Research Question 3 How do group differences in linguistic/emoji knowledge affect how signal conflict is resolved?

4.2 Norming study

Norming studies are a standard practice in any experimental research that aims to understand human behaviour (Cardillo *et al.*, 2010; Canessa *et al.*, 2021). They are used to evaluate the characteristics of potential stimuli, in order to ensure that a participant's response to a particular stimulus can be attributed more directly to a specific property of that stimulus. It may be tempting to assume that a particular image, word, sentence or sound has the specific properties of interest purely because the researcher considers this to be the case. This may result in stimuli which only elicit a response in participants who are similar to the researcher (e.g. age, sex, ethnicity). It also makes it difficult to create a baseline for comparing effects, as stimuli which the researcher considers neutral may be

anything but to some participants[2]. Better practice is to start with a very large set of potential stimuli, fully test their specific properties with people of relevant characteristics, then select only those items which are the most prototypical and least ambiguous examples.

As noted above, both the delineation of social categories and the interpretation of cues to those categories depend heavily on cultural context. Therefore, any study connecting the two must be careful to choose participants who are likely to share a common cultural context, and thus a shared interpretation of the relevant social categories and how they might be communicated. To address this issue, the main experiment recruited participants from an online platform (Prolific) whose profile listed them as native English-speakers living in London, with ethnicity listed as either Black or White. The platform asks "What ethnic group do you belong to?" with possible responses being Asian, Black, Mixed, White, or Other.) I also use the terms Black and White as labels for the identities that participants infer in the experiment, based either on Tweet text alone, or text with emoji.

In order to present participants in the main study with tweets that would have real characteristics of Black or White authors from the same cultural context as the participants, I selected the stimuli using a norming study.

From users in the Users Dataset (Section 3.2) annotated as having either lighter (T1/T2) or darker (T3/T5) skin, I selected 50 tweets each, re-sampling until tweets contained no profanity, sensitive topics or personal information, while preserving the ratio of darker-skinned to lighter-skinned authors. While not all of these authors would necessarily present or identify as Black or White, I rely on the norming study to select tweets from this set that *do* index these identities. All tweets ended with a TME (e.g. 👍🏿, 👍🏾, 👍🏽, 👍🏼, 👍🏻 or 👍) and contained no

[2]When I was a research assistant in psycholinguistics, I was tasked with generating lists of words to be used in an experiment measuring the effect of emotional affect on processing times. This required words classified as negative, neutral and positive. The norming study revealed that 'cow', while predictably rated neutral by online participants in Europe, was considered very positive by participants in India.

other emoji.

To determine the extent to which the text alone is informative about author identity, I use a modified version of the Matched Guise Test as introduced in Chapter 2, following Staum Casasanto (2010). The reader is shown a tweet with the emoji removed. Rather than complete a multi-point questionnaire as in the standard Matched Guise Test (Lambert *et al.*, 1960), the reader is forced to make a simple binary choice --- is the author of the text Black or White? There is no option for the reader to say they are unsure or to skip a stimulus. When unable to decide, they must pick at random. The outcome is the ratio of Black:White judgments for each stimuli. This ratio will be approximately 50:50 if participants as a group do not make any association between the linguistic properties of the text and author identity. However, if there is such an association, this will skew the ratio of Black:White judgments in the appropriate direction.

The use of a forced binary choice, with no possibility to express uncertainty, is widely used in research into human decision making processes and is better known in experimental psychology as the two-alternative forced-choice task. In that context, the method is often used to understand decision making processes, with a dependent variable such as accuracy or reaction time (Ratcliff and Rouder, 1998; Bogacz *et al.*, 2006). However, the use of a binary choice with no option to express neutrality or uncertainty also makes it possible to detect small effects when results are aggregated over participants. This is especially important for effects that may be subconscious, such as attitudes towards, and preconceptions of, language or identity as in both the norming and main study, where I use the same forced choice methodology.

Ethics approval was obtained from the University. Each consenting participant was shown all 100 texts, in randomised order. 40 participants in total were recruited through the Prolific platform. Prolific allows researchers to recruit participants based on a wide range of voluntarily-declared demographics that they

Figure 4.1: Results of the text norming study. Each bar is a text, comparing the ratio of Black (dark brown) to White (light brown) judgments. Rows 1 and 2 show the ratio of Black:White judgments for all four combinations of author/reader ethnicity. Row 3 shows the mean ratio of each text, combining Black and White reader judgments. Texts are ordered by mean ratio. Standard deviation bars denote the extent to which the two reader groups agree in their judgments for a given text. Red line denotes the point of at-chance judgment.

provide in their profile. I recruited 20 participants who self-identified as Black and 20 who identified as White. The outcome was 40 ratings per text, from which the ratio of Black:White judgments for each was calculated. Participants were paid £2.25 (hourly rate: £8.20).

The results for each of the 100 texts are shown in Figure 4.1. Many are judged similarly by both Black and White participants, but the standard deviation for the proportion of White judgments for some texts is as high as 0.32. Mean standard deviation for all stimuli is 0.09 (σ=0.07).

To classify texts as being Black-aligned, non-aligned or White-aligned I used a simple heuristic based on the level of agreement between Black and White participants. Using the aggregate ratio over all participants would introduce a confound in the experiment --- each stimuli should be perceived the same way by all participants. Therefore, Black-aligned/White-aligned texts are those where over

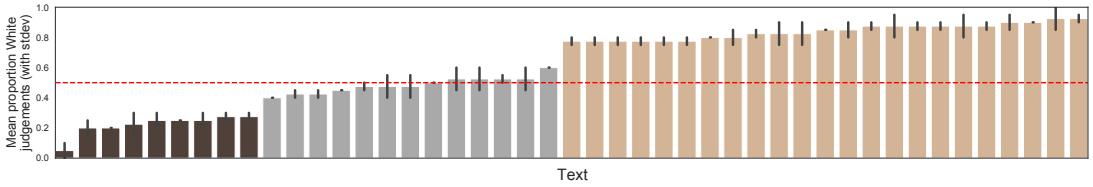


Figure 4.2: Mean proportion of White authorship judgments of final stimuli meeting selection criteria, with standard deviations showing variation between judgments of Black and White readers. Red line denotes the point of at-chance judgment. Bar colour denotes classification of stimuli based on proportion of White judgments by all readers: Black-aligned text is ≤ 0.3, non-aligned text is between 0.4 and 0.6, White-aligned text is ≥ 0.7.

70% of *both* Black and White participants rated the text as having a Black/White author. Non-aligned texts are those rated as having Black authorship by between 40% and 60% of *both* Black and White Participants. 45 tweets did not meet any of these criteria and were discarded. This left 9 Black-aligned texts, 13 non-aligned texts and 23 White-aligned texts. Figure 4.2 shows the mean ratio of judgments per text by all participants. The mean standard deviation for these texts is 0.05 (σ=0.04), with a maximum of 0.11. This subset of stimuli therefore exhibit a higher degree of agreement between Black and White participants than the whole set.

4.3 Experimental setup

From the norming study, I randomly selected 8 Black-aligned, non-aligned and White-aligned texts, for a total of 24 texts. These are shown Table 4.2, with a single example laid out in Table 4.1. Each text was manipulated to create 4 critical stimuli differing only in terms of their emoji content for a total of 96 stimuli --- one with no emoji, one with a Black-aligned emoji, one with a White-aligned emoji, one with a yellow emoji. These 96 stimuli were arranged into 4 blocks of 24. Each

Use	Details	Example
Source data	The original Tweet	Next step manager then owner. By 28 just watch me 🤷🏿‍♀️
Norming study	Used to determine linguistic bias	Next step manager then owner. By 28 just watch me
Main study	Critical stimuli, no emoji	Next step manager then owner. By 28 just watch me
Main study	Critical stimuli, Black-aligned emoji	Next step manager then owner. By 28 just watch me 🤷🏿‍♀️
Main study	Critical stimuli, Yellow emoji	Next step manager then owner. By 28 just watch me 🤷‍♀️
Main study	Critical stimuli, White-aligned emoji	Next step manager then owner. By 28 just watch me 🤷🏻‍♀️

Table 4.1: Example of stimuli generated from a single Tweet, written by a Black author, ending with a tone-modifiable emoji. The text alone was considered linguistically non-aligned (approximately equal numbers of Black and White judgments) by participants of the norming study.

block contained each sentence only once, with each of the four emoji manipulations appearing six times. Participants were assigned to a block at random. These stimuli allow us to measure a response variable (whether a participant selects Black or White) in relation to one of three explanatory variables: reader ethnicity (Black or White), linguistic bias of the text (Black-aligned, non-aligned, White-aligned), or the emoji content of the text (none, Black-aligned, yellow or White-aligned).

In order to obfuscate the focus of the experiment (risking participants simply matching their response to the emoji), 48 filler trials were included. The filler trials asked participants to choose between a female/male or an old/young author identity, rather than a Black/White identity. The text was provided by random tweets (containing no profanity, sensitive topics or personal information) from the Users Dataset. The four emoji types were also balanced across the fillers to prevent the critical trials being conspicuous due to emoji only appearing when the choice was between Black or White.

Each participant therefore completed 72 trials (48 filler, 24 critical) --- 6 critical trials with no emoji, 6 with Black-aligned emoji, 6 with White-aligned emoji and 6 with yellow emoji. These stimuli were presented to participants using the same

modified Match Guise Test (Staum Casasanto, 2010) as the norming study, again in randomised order. The only difference is that participants were presented with a binary response question about the age, sex or ethnicity of the author (depending on whether the trial was critical or filler), and stimuli could also contain emoji. As with the norming study, there was no option for participants to say they are unsure or to skip a trial and they are instructed to select at random when they are unable to decide. The outcome is the ratio of Black:White judgments for each stimuli. As per Staum Casasanto (2010), the ratio will be approximately 50:50 if participants *as a group* do not make any association between the linguistic properties of the text and author identity. However, if there is such an association, this will skew the ratio of Black:White judgments in the appropriate direction.

As in the norming study, I obtained ethics approval from the University. I recruited 488 participants via the Prolific platform, based on their self-reported demographic details. None had taken part in the norming study. All were native English speakers and based in London, to maximise familiarity with the linguistic content of the stimuli. Half the participants self-identified as Black, half as White, as determined by demographic information provided by Prolific. This number of participants was selected based on a power analysis conducted with the G*Power software package (Faul *et al.*, 2007). This sample size gives the statistical analysis (using Fisher's Exact Test) a power of 0.9 to detect an effect size of 0.15 (that is, the difference of proportions between two measures being tested) with the standard 0.05 alpha error probability.

Being able to detect a smaller effect size or achieve greater power would require between 500 and 700 participants in each group. This was not feasible, given the low numbers of Black participants registered on the Prolific platform. I therefore aimed for the greatest power/smallest effect size possible. For statistical analysis, to determine whether there is a significant difference in the the ratios of judgments for stimuli based on factors such as emoji type, text type or participant ethnicity,

I use Fisher's Exact Test and report odds ratios and p-values.

Participants who indicated they may not have been paying due attention by completing the experiment too slowly or too quickly (± 2 standard deviations from the mean completion time of 9.9 minutes (σ=8.8) for Black participants, 6.7 minutes (σ=2.3) for White), or who simply pressed the same response key on every trial, were replaced with new participants prior to any analysis being performed. This amounted to 14 participants excluded for taking too long: 6 Black and 8 White. Excluded participants were still paid the standard rate for their time: £1.56 for a ten minute experiment (hourly rate: £9.36).

4.4 Results and analysis

In all cases, I generate a 2x2 contingency table for the response variable (whether the reader selected Black or White) with respect to one relevant explanatory variable (reader ethnicity, text type, emoji type), with the other two held constant. For example, analysing the effect of Black-aligned and White-aligned text on responses would generate a 2x2 table for each reader ethnicity group. Rows represent Black and White judgments, the columns represent the two text alignments. Therefore, cells report the number of each judgment type for each of the text alignments. Analysing this table using Fisher's Exact Test determines if there is a statistically significant difference in the proportions of the values of the response variable, relative to each value of the explanatory variable.

4.4.1 Baselines

Baseline results are presented in Figure 4.3. These are reader judgments of Black-aligned/non-aligned/White-aligned texts containing no emoji. The non-aligned texts are judged at chance by both participant groups (Black: μ=48.77%; White: μ=51.23%), while the ratio of judgments for Black-aligned texts (Black:

ID	Text	Bias	Black	Yellow	White
1	Big tune after Big tune	Black	🙌🏾	🙌	🙌🏻
2	Finally got that fresh cut	Black	👌🏾	👌	👌🏻
3	I dont sip the syrup, i got friends to lean on	Black	🤙🏾	🤙	🤙🏻
4	If you don't feel like your girl a 10 then idk	Black	🤷🏿‍♀️	🤷‍♀️	🤷🏻‍♀️
5	My ancestors definitely gave me some guidance this year boy	Black	🙏🏾	🙏	🙏🏻
6	New month, New blessings, New beginnings	Black	✌🏾	✌	✌🏻
7	Shout out to everyone that's fasting and had to use the central line today	Black	✌🏾	✌	✌🏻
8	Wish I had the money to dip this country	Black	🏃🏿‍♂️	🏃‍♂️	🏃🏻‍♂️
9	Creed is a great film	Neutral	👏🏾	👏	👏🏻
10	Feeling so brand new right now	Neutral	💅🏾	💅	💅🏻
11	Fool me once, shame on you. Fool me twice, shame on me	Neutral	🤷🏿‍♀️	🤷‍♀️	🤷🏻‍♀️
12	Got me feeling like i can get any polo	Neutral	🙌🏾	🙌	🙌🏻
13	Looking out the window was my tune	Neutral	🙌🏾	🙌	🙌🏻
14	Next step manager then owner. By 28 just watch me	Neutral	👩🏾‍🦱	👩‍🦱	👩🏻‍🦱
15	No matter the path you take, you're going to have to work hard. So you might as well work hard for what you're really passionate about	Neutral	🤷🏿‍♂️	🤷‍♂️	🤷🏻‍♂️
16	Stay positive. Stay focused. Stop worrying what could go wrong. Instead, focus on what can go right. Be a warrior, not a worrier	Neutral	👍🏾	👍	👍🏻
17	Does blowing on tea actually make it cooler or is it just fun to make the waves in your cup?	White	🤷🏿‍♀️	🤷‍♀️	🤷🏻‍♀️
18	Fourth consecutive weekend of long haul flying.	White	🤷🏿‍♀️	🤷‍♀️	🤷🏻‍♀️
19	I am ageing a year every minute	White	🧔🏾	🧓	🧓🏻
20	Need to get out! What's everyone doing tonight?	White	🤷🏿‍♂️	🤷‍♂️	🤷🏻‍♂️
21	Never too old for Christmas	White	👌🏾	👌	👌🏻
22	Premiership games on a Friday the start of something very special	White	🙌🏾	🙌	🙌🏻
23	The nap that I had was perfect	White	👌🏾	👌	👌🏻
24	Today's spirit animal is the man on my train who has commandeered a six seater area to lie down and go to sleep	White	👏🏾	👏	👏🏻

Table 4.2: The 24 text stimuli used in all experiments, with the three emoji used for each one. The text's linguistic bias, as determined through norming, is also shown.

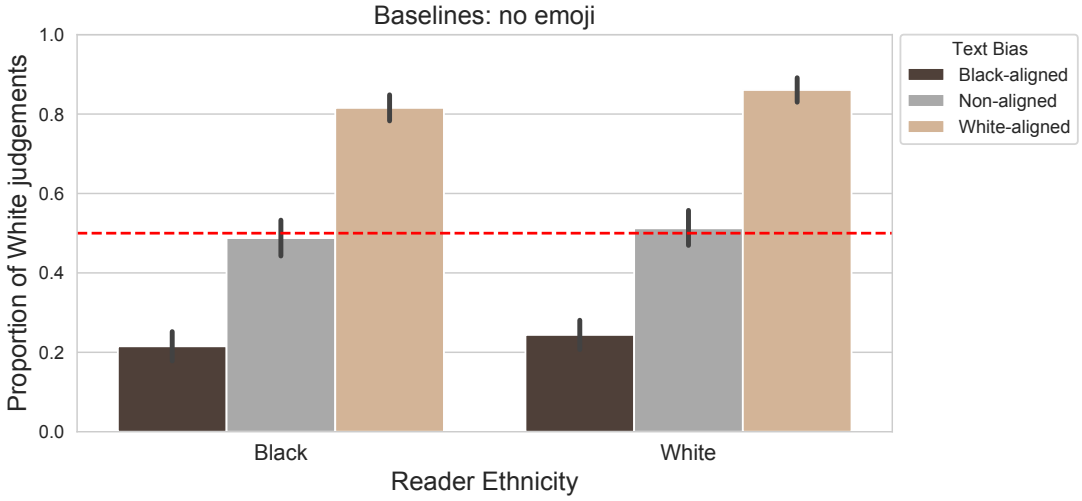


Figure 4.3: Baseline results: Proportion of White judgments (with 95% CI) for different text types, with no emoji added. Red bar denotes judgments at chance.

μ=21.52%; White: 2μ=4.39%) and White-aligned texts (Black: μ=81.56%; White: μ=86.07%) is skewed in favour of the text type.

4.4.2 H1 and H2: How readers interpret skin-toned emoji

Hypothesis 1 tests whether readers perceive skin-toned emoji as salient indicators of author identity, while Hypothesis 2 tests whether this perception is different in Black and White reader groups. Figure 4.4 shows results for both hypotheses. The ratio of judgments by Black readers for non-aligned texts containing emoji is significantly different from their baseline (Fisher's exact test. Black-aligned emoji: odds ratio=2.887, p=9.0693e-15; White-aligned emoji: odds ratio=1.691, p=6.4329e-05). White readers are also significantly different from their baseline (Fisher's exact test. Black-aligned emoji: odds ratio=2.413 p=3.9385e-11; White-aligned emoji: odds ratio=1.934, p=7.0562e-07). Comparing Black and White reader groups reveals no significant differences for Black-aligned nor White-aligned emoji (Fisher's exact test. Black-aligned emoji: odds ratio=1.32, p=0.062412; White-aligned emoji: odds ratio=1.262, p=0.094715).

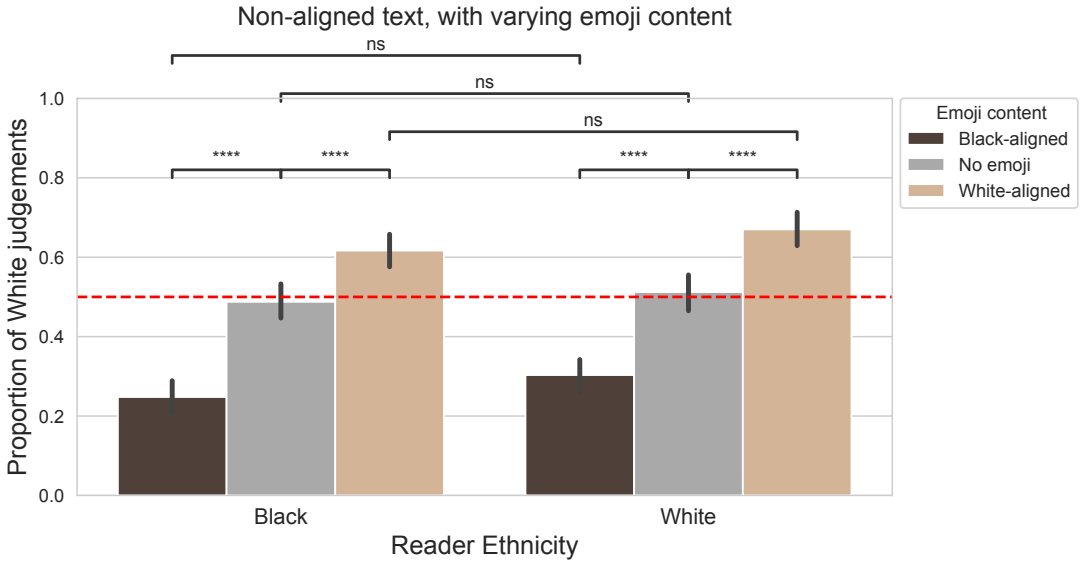


Figure 4.4: Proportion of White judgments (with 95% CI) for non-aligned texts with Black-aligned or White-aligned emoji. Non-aligned texts with no emoji are shown, as baseline reference. H1 compares within participant groups; H2 between participant groups. Red bar denotes judgments at chance. Brackets denote the interval that p-value falls within: ns: (0.05, 1]; *: (0.01, 0.05]; **: (0.001, 0.01]; ***: (0.0001, 0.001]; ****: $p \leq 0.0001$.

Accordingly, the data support Hypothesis 1 but not Hypothesis 2 --- whilst readers do use Black-aligned and White-aligned emoji to determine authorship, no evidence was found to support a claim that Black and White reader groups use emoji in this way to different extents.

This finding may seem at odds with the results of Section 3.2.1, which found that London Twitter users with darker skin tones are more likely to use TME+. However, differences in perception are not guaranteed where differences in production are found. The norming study shows that readers in both the Black and White user groups are able to perceive both White and Black linguistic signals. Emoji as identity props, as per Goffman, can be asymmetric in nature under Jenkins: the identification process which enables me to understand your identity through your linguistic habits does not require me to have those same habits, only to recognise them and attach some social meaning. From this perspective, Hypothesis 2 was poorly constructed as it conflated production with perception.

A further analysis (not included in the pre-registration) suggests that although there is no evidence of difference in how Black and White reader groups perceive identity through emoji, there is a difference in how readers in general perceived the actual Black-aligned and White-aligned emoji. Specifically, when aggregating both reader groups (since there is little difference between them), adding a Black-aligned emoji to neutral text pushes the proportion of Black judgments from approximately 50% to, on average, 72.44% (σ=3.91). For a White-aligned emoji, White judgments make up 64.34% (σ=3.77) of responses. The deviation from random chance is not symmetric for both emoji: on average, a Black-aligned emoji elicits a stronger response of 23%, compared to 14% for a White-aligned emoji.

One account for this disparity is that Black-aligned emoji are more strongly associated with authentic expression of identity, exactly because they are produced for that reason (Robertson *et al.*, 2018, 2020). An alternative account is that Black-aligned emoji, which were presented on a light background, are just visually more

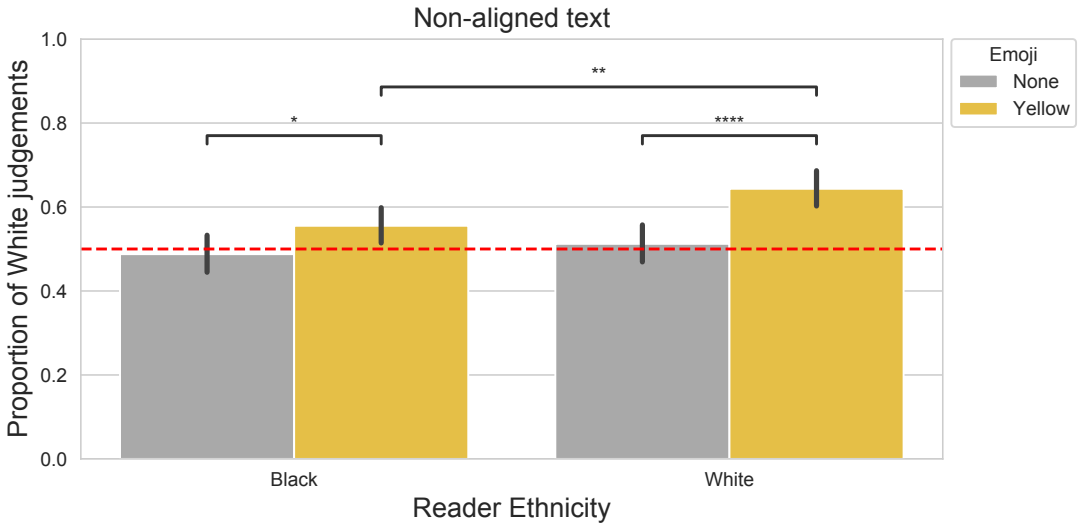


Figure 4.5: Proportion of White judgments (with 95% CI) for non-aligned texts with yellow or no emoji. Red bar denotes judgments at chance. Brackets denote the interval that p-value falls within: ns: (0.05, 1]; *: (0.01, 0.05]; **: (0.001, 0.01]; ***: (0.0001, 0.001]; ****: $p \leq 0.0001$.

prominent in experimental stimuli. However, this is unlikely based on a follow-up study where I measured reaction times to emoji. 58 participants were recruited through social media, receiving gratitude as payment. They were shown 37 text stimuli from the main experiments - 19 fillers with no emoji, 6 neutral stimuli each with a Black-aligned, White-aligned or yellow emoji. Participants pressed a key as quickly as possible if there was an emoji in the text. Trials over 5000ms were discarded. Overall accuracy was 90.5%. Reaction times for correct Black-aligned trials (μ=810ms, σ=475) and White-aligned trials (μ=826ms, σ=532) were not significantly different (Independent samples t-test. t=-0.39, p=0.70).

4.4.3 H3: How readers interpret yellow emoji

Hypothesis 3 tests whether Black readers perceive the yellow emoji as more representing a White than a Black identity, with results shown in Figure 4.5. This proved to be the case, with non-aligned texts with yellow emoji being

perceived significantly differently from the baseline of non-aligned text with no emoji (Fisher's exact test. Black readers: odds ratio=1.31, p=0.04). The same test for White readers found that they also saw such texts as representing a White identity (Fisher's exact test. White readers: odds ratio=1.72, p=4.38e-05). This runs counter to my prediction that only the Black reader group would see yellow as White.

Although RQ3 (discussed later) focused on how Black and White reader groups compare when viewing Black-aligned or White-aligned emoji, I performed a similar analysis here for the yellow emoji[3]. There is a significant difference in how Black and White reader groups perceive authorship of neutral text containing a yellow emoji (Fisher's exact test: odds ratio=1.445, p=0.006). Both rate the yellow emoji as representing a White author, with this effect more pronounced in White readers (+13.11% over baseline) than in Black (+6.76% over baseline). These results confirm that Black readers see the allegedly neutral yellow emoji as representing a White identity. This perception is also held by White readers, and to a greater extent. This difference is not due to a wide margin between baselines: Figure 4.3 shows that these are essentially equivalent across the two reader groups.

Instead, readers may be more likely to perceive yellow as White simply because the yellow and lighter emoji tones are perceptually more similar than the yellow and darker tones. Visual similarity can be quantified by representing the colours within the CIELAB model (Schanda, 2007) and using the Delta E CIE 2000 algorithm (Sharma *et al.*, 2005) to measure their difference. CIELAB was specifically designed such that changes in a colour's values (which measure how black/white, green/red and blue/yellow a colour is) directly correspond to the visual perception of those changes by humans. The larger a Delta E value between two colours, the larger the perceptual difference. Delta E is 58.36 for Black/yellow and 19.38 for White/yellow. Although emoji have only recently become more human-like,

[3]This additional analysis was not explicitly included in the pre-registration document.

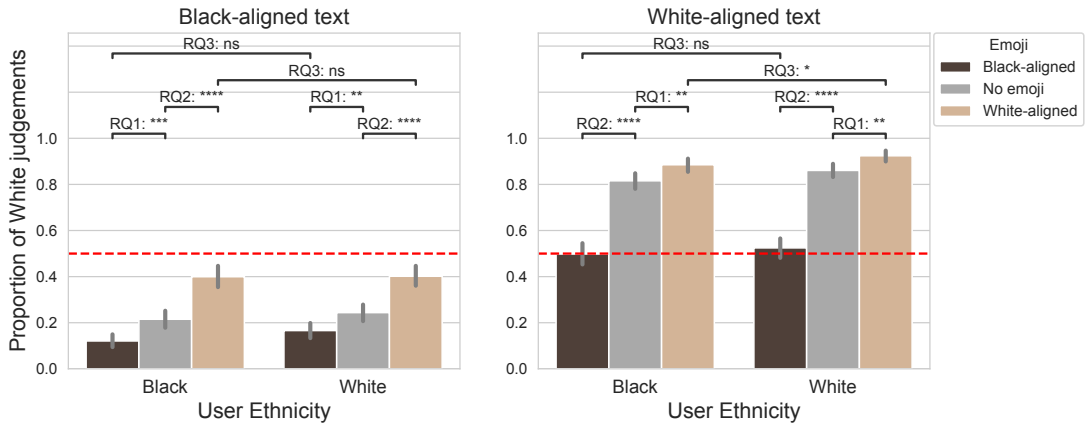


Figure 4.6: Proportion of White judgments (with 95% CI) for Black-aligned text (left) or White-aligned text (right), depending on reader ethnicity and whether a Black-aligned, White-aligned, or no emoji is present. RQ1 and RQ2 test within participants, for congruent and incongruent text/emoji combinations, respectively. RQ3 tests between participant groups for both congruent and incongruent. Red bar denotes judgments at chance. Brackets denote the interval that p-value falls within: ns: (0.05, 1]; *: (0.01, 0.05]; **: (0.001, 0.01]; ***: (0.0001, 0.001]; ****: $p \leq 0.0001$.

it seems that even their cartoonish origins may have (however inadvertently) created a situation of "white first" design (Kayla Heffernan, 2016).

4.4.4 RQ1 and RQ2: The interaction of emoji and language signals

Research questions 1 and 2 examine how different text/emoji combinations affect judgment of author identity. Adding a similarly-aligned emoji to a text (RQ1: Figure 4.6) results in a boosting effect: compared to baseline (text + no emoji) there is a significantly different proportion of judgments in favour of the emoji alignment. For each of the congruent combinations, change over baseline was significant for both Black readers (Fisher's exact test. Black-aligned/Black-aligned: odds

ratio=1.99, p=0.0001; White-aligned/White-aligned: odds ratio=1.74, p=0.00297) and White readers (Fisher's exact test. Black-aligned/Black-aligned: odds ratio=1.62, p=0.0032; White-aligned/White-aligned: odds ratio=1.97, p=0.0018). The effect of emoji alone is slightly lower than that of text alone (as per Figure 4.3 and Figure 4.4), but find a ''boosting effect'' when both signals are present and in agreement. From this I conclude that each signal is distinct and neither one overpowers the other.

When text/emoji alignment are mismatched (RQ2: Figure 4.6) there is instead a dampening effect relative to the text-only baseline: for both reader groups, mismatched text/emoji results in judgments that are significantly different from the text-only baseline in the direction that is closer to random (Black-aligned text, White-aligned emoji - Fisher's exact test. Black readers: odds ratio=2.43, p=5.3505e-10; White readers: odds ratio=2.08, p=1.7704e-07. White-aligned text, Black-aligned emoji - Fisher's exact test. Black readers: odds ratio=4.46, p=6.3499e-26; White readers: odds ratio=5.60, p=8.2273e-31)).

For the particular texts used, the White-aligned text/Black-aligned emoji condition shows both reader groups at chance for judging author identity (the first and fourth bars in the right-hand figure of Figure 4.6), while the opposite arrangement is slightly below chance for both groups (the third and sixth bars in the left-hand figure). This suggests the two conflicting signals do not simply cancel each other out. The final resolution (i.e. how strongly the reader perceives a particular identity) may depends on factor such as what specific linguistic cues are present and how strong they are --- factors which were did not specifically control for in this experiment. As per Section 4.1, I selected texts based only on their *overall* classification by participant groups in the norming study.

Therefore, it could be either that Black-aligned texts in general are a more salient indicator of author identity than White-aligned texts, or simply that the particular texts used had this property.

Careful manipulation of the linguistic content from which text alignment is derived could give a better understanding of what generated this result.

How readers behave prior to making their judgment could also be informative about difficulty of resolving text/emoji incongruity. In psycholinguistics, eye-tracking of participants faced with ambiguity is highly informative as to what extent the reader is aware of conflicts (Rabagliati and Robertson, 2017). Reaction times are commonly used as a proxy for cognitive load: responding to ambiguous words in sentences takes longer (Simpson, 1994). I tentatively report that mean reaction times were reduced when text/emoji were White-aligned/Black-aligned, compared to Black-aligned/White-aligned. Black readers were 2.6s slower (8.7s vs 6.1s), White readers 0.3s (5.0s vs 4.7s). I stress, however, that the experimental paradigm here is not suited to measuring reaction times: I did not control for stimuli length or restrict stimuli presentation time. I offer this analysis only to encourage potential future work on the increased cognitive load that incongruent emoji/text may, or may not, provoke.

4.4.5 Research Question 3: How reader groups compare in perceiving congruent and incongruent signals

Results are again shown in Figure 4.6, marked with RQ3. Comparing results between Black and White readers shows a significant difference in how they judge texts in the White-aligned/White-aligned congruent setting (Fisher's exact test. White-aligned Text/White-aligned Emoji: odds ratio=1.16, p=0.049) but not the Black-aligned/Black-aligned setting (Fisher's exact test. Black-aligned Text/Black-aligned Emoji: odds ratio=1.45, p=0.054). Neither of the incongruent settings show a statistical difference between groups (Fisher's exact test. Black-aligned Text/White-aligned Emoji: odds ratio=1.11, p=0.44; White-aligned Text/Black-aligned Emoji: odds ratio=1.01, p=1.0).

This result is only just past the threshold of statistical significance, with

White readers in this particular condition having a slightly higher proportion of congruent judgments than Black readers (0.924 vs 0.885). Similarly, the result for two Black-aligned signals only just misses the threshold of statistical significance (p=0.055). It is possible that the sample size is too small to reliably detect any small effect that may or may not occur here. If there is a very small but significant difference between Black and White readers in their own congruent conditions, it may be because each group is especially sensitive to a matching text/emoji signal which also matches their own identity, which is triggered by in-group preferences being especially well-catered to in these conditions. However, calculating the Bayes factors for contingency tables (Jamil *et al.*, 2017) to compare Black and White readers in each congruent condition (Bayes Factor: Black-aligned congruent=0.42; White-aligned congruent=0.40) suggests there is anecdotal evidence for *no association* between the explanatory variable (reader ethnicity) and the response variable (their judgment of the author's identity).

4.5 Discussion

When participants read an otherwise neutral text, the presence of a TME was enough to sway them (on average) toward ascribing a particular ethnic identity to the author: Black if the emoji had a dark skin tone, and White if it had either a light skin tone or no tone at all (yellow). Moreover, when the text contained other socially meaningful information about the author's ethnic identity, the effect of this information was moderated by the emoji, indicating that text and emoji provide complementary sources of information.

This study was necessarily limited in scope, with findings derived from a single set of participants within a particular region. Moreover, use of the labels Black and White (as well as the grouped design of the study) glosses over many

possible differences between participants in terms of how they understand these terms with respect to themselves and others. Therefore, even within London, some individuals may not demonstrate the same effects as the group average observed, and it is not known the extent to which groups from other regions would show the same effects for the toned emoji (especially if the Black/White distinction is less socially meaningful, or differently delineated, in that region). Nevertheless, I have shown that both social media text and emoji *can* be used to infer aspects of identity, and that in this experiment, they *were*. Given the many related sociolinguistic studies indicating that social signalling through language is probably universal, I would argue that even based on just this one study it is likely that social media users who are experienced with emoji are inferring aspects of identity from them all the time. I would also expect that, as found here, emoji and text provide complementary signals. Furthermore, I would expect that for regions of the world where Black and White *are* relevant categories associated with skin colour, qualitatively similar results would be found regarding their association with dark vs light emoji tones; and that in other regions, the emoji skin tones may signal other socially relevant categories. Follow-up work should test these predictions, as well as investigate potential individual differences.

Although I found no between-group differences in the perception of ethnic identity from emoji, such differences might still exist between other groups or for other aspects of identity. Existing work has shown that readers of different ages and genders differ in their interpretations of the pragmatic (Herring and Dainas, 2020) and emotional (Weiß *et al.*, 2020) functions of emoji --- i.e., the author's intentions. It seems likely that there can also be differences in the interpretation of social meaning, especially since on social media, authors and readers may share much less social context than speakers and listeners usually do. Therefore, an important future step would be to consider what happens when the social signal in an emoji is perceived in ways that differ from its intended meaning.

The experiment also addressed the question of whether the default yellow emoji is in fact ‘‘neutral’’. Among the participant groups, I found that it was not. I suggested one possible reason is that the yellow colour is more visually similar to the lighter tones associated with the White identity than to the darker tones associated with the Black identity. However, it is also possible that the association with White has less to do with visual similarity than with the fact that the yellow emoji is the default. Within the British socio-cultural context, where White is the historically dominant and ‘‘default’’ category and other ethnicities are seen in contrast to it (see work on the experiences of migrants to the UK and what it means to be White in Britain (Hickman *et al.*, 2005; Fox *et al.*, 2015; Halvorsrud, 2019)), the yellow emoji may be associated with this identity precisely because of its default status. In that case, in other parts of the world with different demographics or different social hierarchies, the yellow emoji may be associated with other groups.

A related issue is whether the perceived Whiteness of the yellow emoji is a *consequence* of the introduction of the other skin tones, or whether it already had that connotation beforehand. This question is difficult to answer in retrospect, but since this study highlights a clear issue that could arise with other self-representations (a supposedly neutral representation that either isn't, or transitions away from neutrality following other changes), designers or researchers could follow up by studying some future change both before and after it occurs, to discover whether and how user perceptions change as a result of new options. While it is clearly important for users to feel they have options for self-representation, these results suggest that it may be difficult to provide such options without inadvertently advantaging one group over another. Nevertheless, considering these issues in advance can at least reduce the chances of perpetuating ‘‘white first’’ design (Kayla Heffernan, 2016). I recommend designers consider these empirical findings within the context of Critical Race Theory for Human Computer

Interaction (Ogbonnaya-Ogburu *et al.*, 2020).

In addition to these implications for research and design, the above findings also have broader social implications. The inferences that a reader makes about an author's identity may change how the reader feels or acts toward the author, or the author's content. This can have both positive and negative ramifications. A recent example is found in a study by Fitzpatrick (1988), of players in a trust-based investment game who were permitted to communicate either through plain text or text with emoji. Emoji users saw greater financial gains from partners, compared to non-users. And although the author set out to explore the impact of emoji usage in general, the results showed evidence that usage of gendered/skin-toned emoji particularly impacted trust in participants. In a more sinister vein, there is recent evidence (Freelon *et al.*, 2020) of ''troll farms'' engaging in a form of ''digital blackface'' (Princewill, 2017) in order to further their agenda by spreading disinformation through accounts made to look like Black activists. Freelon *et al.* (2020) measured digital blackface through manual assessment of screen names, profile descriptions and tweet content. Given the finding that people use emoji to index identity and social categories, an open question now is whether this can be used to manipulate readers in the same way that profile photos and screen names have been.

Finally, there are implications for end users. Using skin-toned emoji on social media provides a strong signal from which one's ethnicity (if not other social categories) may be inferred --- especially if it matches other linguistic signals. Twitter users may already be sensitive to this, as seen by the phenomenon of tone modulation whereby users deliberately turn skin tones off for specific messages (Section 3.3.2). Although that analysis considered users who already displayed a profile photo of themselves, the same study also showed that users without a profile photo used skin-toned emoji to the same extent as those with a public photo on display. For all of these users, simply determining the most common

emoji skin tone used in their online messages would reveal their ethnicity --- a possibility which may alarm users who deliberately avoid using a profile photo.

This work demonstrates yet another example of how social media data may provide information about users that they have not declared directly --- for example, a user's stance on particular topics can be inferred from network analysis (Aldayel and Magdy, 2019), public tweets aimed at private accounts can result in leakage of potentially personal information (Keküllüoglu *et al.*, 2020), and a user's views on certain political issues are associated with particular linguistic choices, even when not discussing that issue (Shoemark *et al.*, 2017b). Users would quite rightly be disturbed if Twitter had a ''search by skin colour'' function, but emoji arguably facilitate the determination of exactly this information. One option is for platforms to disallow search for skin-toned emoji or to include *all* toned variants in search results, to prevent identification and targeting of specific groups of users.

4.5.1 Update: statistical analysis procedure

The p-values reported above were not corrected for multiple comparisons. Following the advice of my thesis examiners, I have applied a Bonferroni correction for the 23 statistical comparisons undertaken in this chapter. This is the simplest and most conservative approach --- it adjusts the alpha value (the point below which Fisher's Exact Test is considered significant) for the entire set of comparisons by dividing the original alpha. For these experiments, where the original value for alpha was 0.05, the new value becomes 0.05 divided by 23, or approximately 0.002.

This changes only the outcome for a single result in RQ3 - the barely-significant value for the White-aligned Text/White-aligned Emoji condition, which is now considered insignificant along with the three other results of that analysis.

Rather than entirely rewrite the preceding section, I instead add the above note. I feel this is appropriate as the preceding text does include an alternative

Bayesian analysis which did not suffer from this particular methodological error. The corrected frequentist analysis matches that of the Bayesian approach and from this point in the thesis, all statistical analysis is Bayesian.

4.6 Chapter Summary

The main findings, based on a set of controlled behavioural experiments (n=488) which were pre-registered with the Open Science Framework[4] are as follows:

1 Skin-toned emoji provide a salient signal of author identity and readers readily perceive this.

2 This signal is complementary to any encoded in the other linguistic content of a social media text, meaning that the appropriate emoji can either boost an existing signal (if signals are congruent) or dampen it (if not).

3 There is no evidence of differences in how Black and White readers perceive Black-aligned and White-aligned emoji, despite asymmetry in how Black and White authors (in general, and in London specifically) produce these emoji on social media (Section 3.2.1).

4 The yellow emoji is not perceived as neutral by either Black or White readers (contra Pardes (2015); Dewey (2015)). On average, both groups perceive it as more likely to index a White identity, and this effect is stronger in White readers.

This confirms that, in terms of their ability to communicate social meaning, emoji can act in many ways like language. Hopefully this will inspire further work aiming to generalise these findings to groups from other cultural or linguistic

[4]Available online at https://osf.io/tn8mg/

backgrounds, draw out subtler effects, and explore the consequences of readers' inferences on their social media behaviour.

That yellow emoji are not viewed as neutral, and also carry social meaning, is important for designers considering how to implement systems that offer equal opportunities for user representation. This supports claims that even when a supposedly neutral technological option is provided, it may unwittingly build in (or take on) assumptions that create an unbalanced system which better suits one group of users over others. This study also shows how claims of neutrality can, at least in some cases, be tested experimentally with the Matched Guise methodology, and I hope that this will inspire designers to consider similar approaches at earlier stages of the design process.

With evidence that emoji can play a specific role in the identification process, the next chapter will explore whether the connection between emoji and identity can also, like language, influence the behaviour towards those associated with that identity.

Chapter 5

Behaviour

This chapter is based on work published in the paper "Identity Signals in Emoji Do not Influence Perception of Factual Truth on Twitter" (Robertson et al.*, 2021b).*

Chapter 4 demonstrated experimentally that readers consistently associate particular TME+ with specific identities. This outcome parallels the ability of specific linguistic characteristics of speech to index social information about speaker identity (Section 2.4). When a listener connects a speaker with a particular identity on the basis of linguistic signals, the listener's attitude and behaviour towards the speaker can be affected (Section 2.5) even if there is no explicit awareness of the linguistic signal.

In this chapter, I test hypotheses on the possibility that TME+ can influence a reader's behaviour. The experimental procedure, hypotheses and analysis plan were all pre-registered with the Open Science Foundation[1].

5.1 Hypotheses

Hypothesis 1 Readers will be sensitive to the identity signal encoded in emoji and in photos.

[1]Available at https://osf.io/a8r6q

Hypothesis 2 Readers will exhibit a preference for one identity encoded in emoji or in photos.

Hypothesis 3 The combined effect of emoji and photo together will be greater than that of emoji or photo alone.

Hypothesis 4 Yellow emoji encode a White identity. White participants will react to yellow emoji in the same way as White-aligned emoji.

First, I predict that the identity signal encoded in emoji will have an effect on how readers evaluate trivia statements in terms of being true or false, similar to the findings of Lev-Ari and Keysar (2010). I also assume that a profile photo also encodes an identity signal as per Ge *et al.* (2016).

Second, readers will be sensitive to the kind of identity encoded in emoji and photos. Readers may prefer their own identity (similar to in-group bias in favour of own gender (Rudman and Goodwin, 2004)) or they may prefer another identity, for example if all users show a similar bias against a particular group (Valla *et al.*, 2018) or if there are widely-held social stereotypes in favour of that group, as was found by Lambert *et al.* (1960).

Third, the presence of both a photo and an emoji signal will have a larger effect than either one alone, as per the boosting effect observed in Chapter 4 of an emoji and a language signal of identity when users were asked to guess the identity of the author.

Finally, give the finding that White-aligned emoji encode a White identity for White readers (Section 4.4.3) I predict that the same pattern of behaviour will be observed for White readers when presented with either yellow or White-aligned emoji.

These hypotheses are tested separately with self-identified Black and White participants, to ensure I do not mistakenly extend any effects found in one group to all groups, as there are differences in how Black and White social media users use

(Chapter 3) and perceive (Chapter 3) skin-toned emoji in relation to expressions of identity.

5.2 Experimental Setup

The experimental setup is based on Lev-Ari and Keysar (2010), who looked at the effect of a speaker's accent on how listeners react to what they hear. Participants listened to speakers with native, mild foreign or strong foreign accents reading trivia facts such as ''Ants never sleep''. On a scale from 0 to 100, participants rated statements from *definitely false* to *definitely true*. Statements expressed by accented voices were rated as less true, even though participants were informed that the speakers were simply reciting statements written by the native-speaking experimenters. Lev-Ari and Keysar focused on processing difficulty as the explanatory factor for this outcome, rather than out-group or ethnic bias, but the foreign accents used in the experiments are likely a highly salient indicator of ethnicity. To emulate the ''reciting'' aspect, I use retweets.

The experiment closely follows the setup of Lev-Ari and Keysar (2010) with the main difference that it was conducted online. Participants were shown 32 tweets containing trivia and informed that the trivia could be true or false. Their task was to mark on a slider from 0 to 100 whether they thought the trivia was *definitely false* or *definitely true*. Participants were explicitly asked not to look up any trivia and informed that the correct answers would be shown upon completion of the study.

5.2.1 Stimuli

Critical trials manipulated the identity signals present in emoji and photos. Emoji were either dark-toned (aligned with Black identities), light-toned (aligned with White identities) or absent. Only the downwards pointing finger, 👇🏿 👇🏻, was

used. Photos signalled a probable Black identity, White identity or no obvious identity. There were 7 critical combinations of emoji/photo. Photos were female faces generated by StyleGAN[2], to avoid using photos of real people, and blurred to obscure all detail except skin tone. In a pre-norming study run through the Prolific platform, 25 Black and 25 White participants guessed the ethnicity of the blurred faces. I retained only those photos which the majority ($\geq$95%) of participants considered to show a Black or a White person.

All trivia in the critical trials were true, but were likely to be neither well-known to be true nor generally thought (mistakenly) to be false. This was to prevent any ceiling effect in the ratings of the stimuli and was determined through a pre-norming study of trivia facts taken from Snapple soft drink lids[3]. 20 Black and 20 White participants rated 100 random facts as either true or false. I randomly sampled 8 from those facts where participants were at chance in deciding between true and false. Mean ratings for these facts (where 50 is at chance) were 47.5 (σ=0.0932) for Black participants, 44.6 (σ=0.0935) for White participants.

To control for any possible impact of individual facts on participant judgements, each of these facts appeared in each critical combination of emoji/photo. This was done in blocks, such that participants only saw each fact once in a single emoji/photo combination, but across the entire experiment all critical facts were judged in all combinations.

Filler trials (to obscure the intent of the experiment) were balanced across two types of fact. Eight were facts that pre-norming participants generally considered to be true/false (4 of each, with mean rating above 95 or below 5), and twelve were facts I created which were very obviously true or very obviously false (6 of each). In addition, profile photos could be colourful and non-human (e.g. scenery) and blurred ticks, indicating a verified account, could be present next to user names. Also, a wider range of emoji, including skin-toned ones, could appear in

[2]https://github.com/NVlabs/stylegan

[3]See https://www.snapple.com/real-facts for a full list

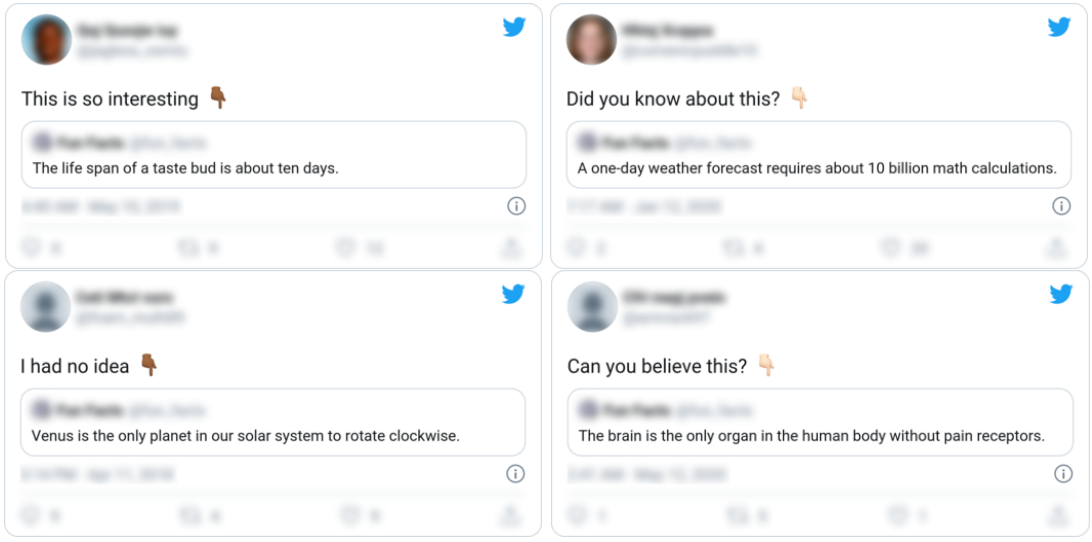


Figure 5.1: Examples of critical trials from one block, showing manipulation of the identity encoded in profile photos and emoji.

the main retweeter's text. In total, participants completed 32 trials: 8 critical and 24 fillers, shown in Table 5.1.

All stimuli were presented as a main user retweeting another account, with some comment expressing surprise or incredulity. Authentic-looking tweets were automatically generated from a copy of the Twitter HTML/CSS and saved as PNG files, to discourage participants looking for answers online by copying/pasting the text. The surprise/incredulity comments were selected based on a pre-norming study. 50 Black and 50 White participants provided three examples of how they would comment if sharing such a tweet. The most common responses shared between Black and White participants were chosen and sampled from at random. Display names, user names, dates, number of likes/retweets/comments were all randomised before being blurred. In critical trials, the account being retweeted was always unverified and had a default grey display photo. In filler trials, these properties were randomly assigned. Finally, trials were presented in randomised order per participant. Example stimuli are shown in Figure 5.1.

Type	Text
Critical	A one-day weather forecast requires about 10 billion math calculations.
Critical	Elephants are afraid of bees.
Critical	The life span of a taste bud is about ten days.
Critical	Every hour more than one billion cells in the body must be replaced.
Critical	Children have more taste buds than adults.
Critical	The brain is the only organ in the human body without pain receptors.
Critical	Venus is the only planet in our solar system to rotate clockwise.
Critical	Most newborns will lose all the hair they are born with in the first six months of life.
Filler (Obv. True)	New York is the biggest city in the USA.
Filler (Obv. True)	The first president of the USA was George Washington.
Filler (Obv. True)	The sun rises in the east and sets in the west.
Filler (Obv. True)	Football is the most popular sport in the USA, followed by baseball.
Filler (Obv. True)	There are over 30 places in the USA named Springfield.
Filler (Obv. True)	The USA is made up of 50 states.
Filler (Obv. True)	The largest US state (by area) is Alaska.
Filler (Obv. True)	People can be extremely allergic to peanuts.
Filler (Norm True)	Greyhounds can reach speeds of 45 miles per hour.
Filler (Norm True)	Marine mammals swim by moving their tails up and down, while fish swim by moving their tails left and right.
Filler (Norm True)	The average human produces 10,000 gallons of saliva in a lifetime.
Filler (Norm True)	You blink over 10,000,000 times a year.
Filler (Obv. False)	Most English people are born with two noses.
Filler (Obv. False)	Dogs have no sense of smell.
Filler (Obv. False)	The earth takes 500 days to travel around the sun.
Filler (Obv. False)	The most western state in the USA is Texas.
Filler (Obv. False)	60% of Canadians have a twin brother or sister.
Filler (Obv. False)	The average person can hold their breath for three hours.
Filler (Obv. False)	The internet was invented in 1776.
Filler (Obv. False)	The Fourth of July is always on a Tuesday.
Filler (Norm False)	A ball of glass will bounce higher than a ball of rubber.
Filler (Norm False)	A bee has five eyelids.
Filler (Norm False)	Seals sleep only one and a half minutes at a time.
Filler (Norm False)	There are more trees on Earth than stars in the galaxy.

Table 5.1: The final 32 stimuli used in the experiment.

5.2.2 Participants

Participants were recruited through Prolific.co on the basis that they are native speakers of English, US citizens resident in the US and self-identified their ethnicity as either Black or White. I did not request or record any non-pertinent personal information (e.g. age, gender), in accordance with the University's data handling protocols. In all pre-norming and experimental tasks, participants were paid at a rate equivalent to the National Living Wage of £9.50 per hour in Scotland. Approval for the study was obtained in advance from the University's ethics board. In total, 944 participants (472 Black, 472 White) were recruited for the main study, to ensure sufficient power to determine with a good degree of confidence whether the null or alternative hypothesis is better supported by the evidence (using a Bayesian statistical analysis, described below). Power was determined using the SSDbain package in R (Fu *et al.*, 2020).

5.3 Analysis Plan

The experiment generates a sequence of true/false ratings (from 0 to 100) per critical combination of emoji/photo. I analyse the means of these ratings for differences using a Bayesian two-tailed independent samples t-test (van Doorn *et al.*, 2020), as implemented by the JASP[4] software package. The Bayesian t-test is parameterised by a prior distribution over assumptions of the effect size. I follow standard practice of using a Cauchy prior with mean 0 and scale 0.707 (Quintana and Williams, 2018). The choice of the Bayesian t-test over the frequentist version is motivated by the results of Section 4.4, which saw borderline p-values. The Bayesian version of the t-test for comparing sample means as it reports Bayes Factors rather than p-values, which makes it possible to quantify the amount of support for or against a hypothesis, rather than simply testing whether the null

[4]https://jasp-stats.org/

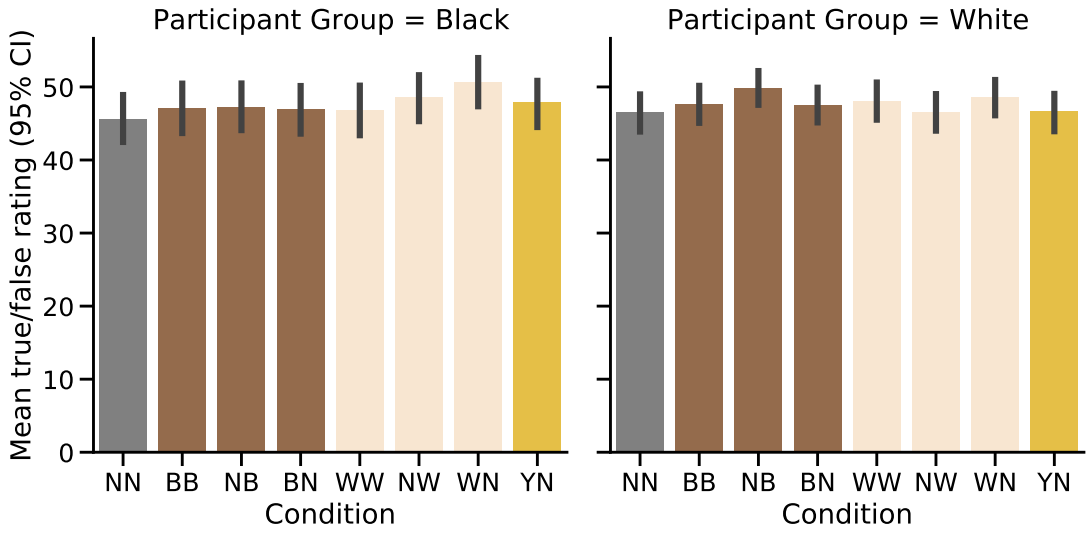


Figure 5.2: Mean true/false ratings (with 95% confidence intervals) per participant group.)

hypothesis can be rejected (Rouder *et al.*, 2009). The Bayes Factor (BF) is the ratio of the marginal likelihoods of two competing hypotheses (H0 and H1), given some observed data. Multiplying the prior odds of H0 and H1 (i.e. the ratio of P(H0) to P(H1)) by the BF gives the posterior odds: the ratio of P(H0|data) to P(H1|data).

The BF quantifies the extent to which the observed data supports the hypotheses under consideration. For this experiment, the two competing hypotheses are no difference in means (H0) and a significant difference in means (H1). A BF value of 1 means the observed data supports each hypothesis equally well (i.e. it is uninformative), and normally indicates additional data is needed or hypotheses need to be reformulated. A BF value greater than 1 means there is more support for H1, while a value less than 1 means more support for H0.

5.4 Results

Mean true/false ratings per experimental condition for each participant group are shown in Figure 5.2. Each bar represents an experimental condition, with

Hyp.	Condition A	Condition B	Description	BF (Black)	BF (White)
H1a	No photo, no emoji	No photo, Black-aligned emoji	Sensitive to presence of emoji	0.109	0.443
H1b	No photo, no emoji	Black photo, no emoji	Sensitive to presence of photos	0.099	0.087
H1c	No photo, no emoji	No photo, White-aligned emoji	Sensitive to presence of emoji	0.186	0.073
H1d	No photo, no emoji	White photo, no emoji	Sensitive to presence of photos	0.834	0.142
H2a	No photo, matching emoji	No photo, non-matching emoji	Sensitive to identity-matching emoji	0.098	0.434
H2b	Matching photo, no emoji	Non-matching photo, no emoji	Sensitive to identity-matching photos	0.284	0.086
H3a	Matching photo, matching emoji	No photo, matching emoji	Emoji encode distinct identity signal	0.085	0.108
H3b	Matching photo, matching emoji	Matching photo, no emoji	Photos encode distinct identity signal	0.084	0.075
H4	No photo, no emoji	No photo, yellow emoji	Yellow emoji encode White identity	N/A	0.073

Table 5.2: Overview of hypotheses tested, in terms of photo/emoji conditions. Except for H4, these were tested separately for Black and White readers: matching/non-matching means that the skin tone of the emoji was aligned with the self-reported ethnicity of the reader. Final columns show Bayes Factors for t-tests and are discussed in the analysis/discussion section.

labels XY representing Emoji status (Black-aligned, White-aligned, No emoji, Yellow) and Photo status (Black, White, No photo). Bars are coloured based on the prevalent identity signal of the emoji.

The results of all analyses for both participant groups were in favour of H0 --- there was no effect on the mean true/false ratings in any subsets of ratings selected for comparison. Table 5.2 reports the results of all t-tests, in terms of Bayes Factor, from which it can be seen that experimental evidence offers between 1.2 and 13.4 times more support (i.e. 1/BF) for H0 in all cases than for the alternative hypothesis of a difference in mean true/false ratings. From this, neither emoji nor profile photo, either alone or together, influenced a reader's behaviour in the true/false trivia task.

5.5 Further Analysis and Discussion

Although the experiment was motivated by a range of prior work, used carefully designed and normed stimuli, and involved almost 1000 participants, the evi-

Figure 5.3: Example of a critical trial from one block, with the user directly tweeting the trivia, rather than retweeting another account.

dence found against any effect of identity signals prompted me to reflect on the experimental design and explore possible issues.

Did the presentation of the fact as a retweet of an account with no identity signals cause readers to focus only on the retweeted account? I ran a small scale version (n=80, White participants only) of the experiment without the retweet (Figure 5.3) and again obtained statistical support only for the null hypothesis (BF range of 0.181-0.783, 1.2 to 5.5 times more support for H0).

Was participants' familiarity with Twitter a factor? The Prolific platform provided data on participants' self-reported Twitter use and I separated groups into Twitter users and non-users. Again, I obtained qualitatively the same results, regardless of whether stimuli were tweets or retweets (BF range of 0.144-0.385, 2.6 to 6.9 times more support for H0).

It may be that profile photos had no effect on behaviour because they were blurred. I had normed these blurred photos to ensure the skin tone was determinable, but this alone may not be sufficient to cause any bias or preferential behaviour in this task. As emoji were not blurred but also had no effect, I suspect it is unlikely that unblurring profile photos would have any effect.

Perhaps I should have included a measure of bias by administering an IAT, as per Stanley *et al.* (2011), and balanced participant numbers not only in terms of

self-reported racial identity but also in terms of pre-existing bias. This I leave for future work and may give an interesting perspective on whether emoji can trigger biases or preferences which are already known to exist, but see prior work on issues with failure to IAT results and poor correlation between IAT scores and behavioural outcomes (Blanton *et al.*, 2009).

The true/false trivia task may itself not be best suited to detecting the effect I set out to find. These facts are by definition trivial, and whether they are true or false is likely unimportant to the reader. By contrast, facts with real-world consequences or emotional valence might be more subject to unconscious biases surrounding the identity of the source. I did not explore that type of fact here, since the experiment would be far more challenging to design, raising both ethical questions (as it would involve spreading misinformation) and design issues (it is difficult to come up with plausible stimuli). However, such an experiment could provide important further evidence regarding situations in which identity does or does not affect perceptions of truth.

Finally, participants may have not been sufficiently motivated. In experimental economics, studies such as those based on the Trust Game offer additional incentives to participants for performance, or include an element of cooperation/competition with a confederate (real or imagined), who can be rewarded or punished through the actions of the participant, which encourages participants to focus on a person.

How any findings from such work would translate to the case of Twitter is not clear. On the positive side, there is strong evidence that the participants in this study were not swayed by identity signals to rate trivia as more or less true/false. This has consequences for work in disinformation. In an analysis of tweets from a Russian government-funded troll farm known as the Internet Research Agency, Freelon *et al.* (2020) found that fake accounts included aspects of racial presentation, using Black profile photos and names associated with Black

Americans. The practice of Black impersonation in this case appeared to result in increased engagement with the content from real Twitter users, in terms of likes and retweets. However, as Freelon *et al.* note, there is no data on the identities of the users engaging with those tweets. One way of reconciling the findings of Freelon *et al.* and this thesis is to suppose that the increased engagement came only from additional fake bot accounts (though not necessarily from the same troll farm, as the authors found little evidence of self-amplification) rather than real users. The real users in my study seem indifferent to such identity signals in the context of behavioural influence. To better understand the interaction of users, identity, behaviour and disinformation, future work in this area should combine the experimental approach of this study and the content perspective of Freelon *et al.* (2020).

5.6 Chapter Summary

A large scale pre-registered experiment failed to detect any differences in how identity signals in messages influence how readers evaluate the truth of those messages. Neither emoji nor profile photos had any effect and this was consistent across both self-identified Black and White participant groups. Reasonable adjustments to the methodology and analysis plan yielded similar results. This null result does have a positive consequence if it can be replicated in follow-up studies: ethnic identity as communicated through emoji or profile photos does not alone automatically affect how content is perceived.

Chapter 6

Conclusion

The specific question of whether emoji have the ability to convey social meaning and to play their own specific role in the expression and perception of identity has been robustly addressed, through both analysis of online data in Chapter 3 and large scale online experiments in Chapter 4. Emoji are not just a tool, in the sense of Goffman, through which we express our identity but a fully understood component of the identity process, rooted in language, as defined by Jenkins.

Evidence for a connection between perception of identity and behavioural outcomes was not found in the experiments of Chapter 5, which can be interpreted in several ways. Either the methodology was not suited to inducing the predicted behaviour or perception of a person's broad ethnic identity is insufficient to influence behaviour in the general public --- or perhaps some mixture of both. As such, this thesis leaves open the door to further work in determining in exactly which situations emoji can influence behaviour --- if any. More broadly there is also the question of which identities are indexed by emoji other than those with skin tones. Given the vast range of identities people hold and the thousands of emoji available, this is likely to require additional exploratory work but the experimental methodologies used here will be readily applicable.

The strong connection between emoji and identity means that both individual

users and platforms may need to deal with new privacy concerns. Even those users who shield their physical identity online, by avoiding the use of profile photos, are at risk of broadcasting their skin colour online. Where emoji are searchable, platforms may be accomplices in identifying and targeting specific groups of users. This is possible for skin tone due to the highly direct mapping between real-life identity and emoji skin tones, but likely for any other emoji with indexical usage for specific groups. Of course, this can be used by users to find others like themselves but could be open to abuse. Platforms may need to provide emoji-specific privacy features for users.

Finally, the wide-spread use of emoji to represent identity also raises questions of how those responsible for the physical design of emoji can provide as many users as possible with accurate representation. Such a goal may be unrealistic, in the face of technical limitations and the scale of human variety, but where innovation fails alternatives should as a bare minimum avoid favouring particular groups over others, either explicitly or accidentally. The methodologies used in this thesis can be used diagnostically to identify where this is the case.

Bibliography

Al Zamal F, Liu W, Ruths D (2012). ''Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors.'' *ICWSM*, **270**, 2012.

Aldayel A, Magdy W (2019). ''Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media.'' *Proc. ACM Hum.-Comput. Interact.*, **3**(CSCW). URL https://doi.org/10.1145/3359307.

Babin JJ (2020). ''Linguistic signaling, emojis, and skin tone in trust games.'' *PLOS ONE*, **15**(6), 1--14.

Barbieri F, Camacho-Collados J (2018). ''How Gender and Skin Tone Modifiers Affect Emoji Semantics in Twitter.'' In ''Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics,'' .

Baugh J (1996). ''Perceptions within a variable paradigm: Black and White racial detection and identification based on speech.'' *Focus on the USA*, pp. 169--182.

Baynes LM (1997). ''If it's not black and white anymore, why does darkness cast a longer discriminatory shadow than lightness - an investigation and analysis of the color hierarchy.'' *Denv. UL Rev.*, **75**, 131.

Beinhoff B (2013). *Perceiving Identity through Accent*. Peter Lang, Bern, Switzerland. ISBN 978-3-0353-0454-1. URL https://www.peterlang.com/view/title/35458.

Berard B (2018). ''I second that emoji: The standards, structures, and social production of emoji.'' *First Monday*.

Berryman C, Ferguson CJ, Negy C (2017). ''Social media use and mental health among young adults.'' *Psychiatric quarterly*, pp. 1--8.

Blagdon G (2013). ''How Emoji Conquered The World.'' URL `https://www.theverge.com/2013/3/4/3966140/how-emoji-conquered-the-world`.

Blakemore E (2019). ''Race and ethnicity, explained.'' *National Geographic*. URL `https://www.nationalgeographic.co.uk/history/2019/02/race-and-ethnicity-explained`.

Blanton H, Jaccard J, Klick J, Mellers B, Mitchell G, Tetlock PE (2009). ''Strong claims and weak evidence: Reassessing the predictive validity of the IAT.'' *Journal of Applied Psychology*.

Blodgett SL, Green L, O'Connor B (2016). ''Demographic dialectal variation in social media: A case study of African-American English.'' *arXiv preprint arXiv:1608.08868*.

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006). ''The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks.'' *Psychological review*.

boyd d (2007). ''Why youth (heart) social network sites: The role of networked publics in teenage social life.'' *MacArthur foundation series on digital learning--Youth, identity, and digital media volume*, **119**, 142.

Bradac JJ, Mulac A, House A (1988). ''Lexical diversity and magnitude of convergent versus divergent style shifting: Perceptual and evaluative consequences.'' *Language & Communication*.

Buchan NR, Croson RT, Solnick S (2008). ''Trust and gender: Behavior and beliefs in the Investment Game.'' *Journal of Economic Behavior & Organization*.

Buchstaller I (2005). ''Putting Perception to the Reality Test: The Case of" Go" and" Like".'' *University of Pennsylvania Working Papers in Linguistics*.

Burge J (2019). ''Correcting the Record on the First Emoji Set.'' URL `https://blog.emojipedia.org/correcting-the-record-on-the-first-emoji-set/`.

Campbell-Kibler K (2009). ''The nature of sociolinguistic perception.'' *Language Variation and Change*, **21**(1), 135.

Canessa E, Chaigneau SE, Lagos R, Medina FA (2021). ''How to carry out conceptual properties norming studies as parameter estimation studies: Lessons from ecology.'' *Behavior Research Methods*, **53**(1), 354--370.

Cardillo ER, Schmidt GL, Kranjec A, Chatterjee A (2010). ''Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor.'' *Behavior research methods*, **42**(3), 651--664.

Chen Z, Lu X, Ai W, Li H, Mei Q, Liu X (2018). ''Through a Gender Lens: Learning Usage Patterns of Emojis from Large-Scale Android Users.'' In ''Proceedings of the 2018 World Wide Web Conference,'' URL `https://doi.org/10.1145/3178876.3186157`.

Child JT, Starcher SC (2016). ''Fuzzy Facebook privacy boundaries: Exploring mediated lurking, vague-booking, and Facebook privacy management.'' *Computers in Human Behavior*, **54**, 483--490.

Coats S (2018). ''Skin Tone Emoji and Sentiment on Twitter.'' In ''Proceedings of the 3rd Digital Humanities in the Nordic Countries Conference,'' .

Davis M, Edberg P (2014). ‘‘Proposed Draft Unicode Technical Report: Unicode Emoji.’’ *Technical Report 51*, Unicode Consortium. URL `https://www.unicode.org/reports/tr51/tr51-1-archive.html`.

Delgado R, Stefancic J (2017). *Critical race theory: An introduction*, volume 20. nYU Press.

Dewey C (2015). ‘‘Are Apple’s new yellow face emoji racist?’’ URL `https://www.washingtonpost.com/news/the-intersect/wp/2015/02/24/are-apples-new-yellow-face-emoji-racist/`.

Dickey M (2017). ‘‘Thoughts on white people using dark-skinned emoji.’’ *TechCrunch*. URL `https://techcrunch.com/2017/10/01/thoughts-on-white-people-using-dark-skinned-emoji/`.

D’Onofrio A (2018). ‘‘Controlled and automatic perceptions of a sociolinguistic marker.’’ *Language Variation and Change*, **30**(2), 261--285.

Farnham SD, Churchill EF (2011). ‘‘Faceted Identity, Faceted Lives: Social and Technical Issues with Being Yourself Online.’’ In ‘‘Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work,’’ URL `https://doi.org/10.1145/1958824.1958880`.

Faul F, Erdfelder E, Lang AG, Buchner A (2007). ‘‘G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences.’’ *Behavior research methods*, **39**(2), 175--191.

Fitzpatrick T (1988). ‘‘The validity and practicality of sun-reactive skin types I through VI.’’ *Archives of Dermatology*. URL `http://dx.doi.org/10.1001/archderm.1988.01670060015008`.

Fox JE, Moroşanu L, Szilassy E (2015). ‘‘Denying discrimination: Status,‘race’,

and the whitening of Britain's new Europeans.'' *Journal of Ethnic and Migration Studies*, **41**(5), 729--748.

Frankfurter Z, Kokoszka K, Newhouse D, Silwal AR, Tian S (2020). ''Measuring Internet Access in Sub-Saharan Africa.'' URL `https://openknowledge.worldbank.org/handle/10986/34302`.

Freelon D, Bossetta M, Wells C, Lukito J, Xia Y, Adams K (2020). ''Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation.'' *Social Science Computer Review*, p. 0894439320914853.

Fu Q, Hoijtink H, Moerbeek M (2020). ''Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor.'' *Behavior Research Methods*, pp. 1--14.

Galloway P (2016). ''The Original NTT DOCOMO Emoji Set Has Been Added to The Museum of Modern Art's Collection.'' URL `https://stories.moma.org/the-original-emoji-set-has-been-added-to-the-museum-of-modern-arts-collection-c6060e141f61`.

Garimella K, Tyson G (2018). ''Whatapp Doc? A First Look at Whatsapp Public Group Data.'' In ''Twelfth International AAAI Conference on Web and Social Media,'' .

Ge J (2019). ''Emoji sequence use in enacting personal identity.'' In ''Companion proceedings of the 2019 world wide web conference,'' pp. 426--438.

Ge Y, Knittel CR, MacKenzie D, Zoepf S (2016). ''Racial and gender discrimination in transportation network companies.'' *National Bureau of Economic Research*.

Goffman E (1956). *The presentation of self in everyday life*. University of Edinburgh Social Sciences Research Centre.

Greenwald AG, Banaji MR, Rudman LA, Farnham SD, Nosek BA, Mellott DS (2002). ''A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept.'' *Psychological review*, **109**(1), 3.

Gu MM, Patkin J (2013). ''Heritage and identity: Ethnic minority students from South Asia in Hong Kong.'' *Linguistics and Education*, **24**(2), 131--141. ISSN 0898-5898. URL https://www.sciencedirect.com/science/article/pii/S089858981300003X.

Haimson O (2018). ''Social Media as Social Transition Machinery.'' *Proceedings of the ACM on Human-Computer Interaction*. URL https://doi.org/10.1145/3274332.

Halvorsrud K (2019). ''The maintenance of white privilege: The case of white South African migrants in the UK.'' *Ethnicities*, **19**(1), 95--116.

Harris CR, Coburn N, Rohrer D, Pashler H (2013). ''Two failures to replicate high-performance-goal priming effects.'' *PloS one*.

Herring SC, Dainas AR (2020). ''Gender and Age Influences on Interpretation of Emoji Functions.'' *Trans. Soc. Comput.*, **3**(2). ISSN 2469-7818. URL https://doi.org/10.1145/3375629.

Hickman MJ, Morgan S, Walter B, Bradley J (2005). ''The limitations of whiteness and the boundaries of Englishness: Second-generation Irish identifications and positionings in multiethnic Britain.'' *Ethnicities*, **5**(2), 160--182. URL https://doi.org/10.1177/1468796805052113.

Hogan B (2010). ''The presentation of self in the age of social media: Distinguishing performances and exhibitions online.'' *Bulletin of Science, Technology & Society*.

Hopper R, Williams F (1973). ''Speech characteristics and employability.'' *Communications Monographs*, **40**(4), 296--302.

Jamil T, Ly A, Morey RD, Love J, Marsman M, Wagenmakers EJ (2017). ''Default "Gunel and Dickey" Bayes factors for contingency tables.'' *Behavior Research Methods*, **49**(2), 638--652.

Jenkins R (2014). *Social identity*. Routledge.

Joblove GH, Greenberg D (1978). ''Color Spaces for Computer Graphics.'' In ''Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques,'' URL `http://doi.acm.org/10.1145/800248.807362`.

Jordán-Conde Z, Mennecke B, Townsend A (2014). ''Late adolescent identity definition and intimate disclosure on Facebook.'' *Computers in Human Behavior*. URL `http://www.sciencedirect.com/science/article/pii/S0747563213002537`.

Kapidzic S, Herring SC (2015). ''Race, gender, and self-presentation in teen profile photographs.'' *New Media & Society*.

Kayla Heffernan (2016). ''Design is as good (or flawed) as the people who make it.'' http://www.uxaustralia.com.au/conferences/uxaustralia-2016/presentation/design-people/.

Keküllüoglu D, Magdy W, Vaniea K (2020). ''Analysing Privacy Leakage of Life Events on Twitter.'' In ''12th ACM Conference on Web Science,'' WebSci '20, p. 287–294. Association for Computing Machinery, New York, NY, USA. ISBN 9781450379892. URL `https://doi.org/10.1145/3394231.3397919`.

Kinzler KD, Corriveau KH, Harris PL (2011). ''Children's selective trust in native-accented speakers.'' *Developmental science*, **14**(1), 106--111.

Kwak H, Lee C, Park H, Moon S (2010). ''What is Twitter, a social network or a news media?'' In ''Proceedings of the 19th international conference on World wide web,'' pp. 591--600. AcM.

Labov W (1972). *Sociolinguistic Patterns*. Conduct and Communication. University of Pennsylvania Press. URL https://books.google.co.uk/books?id=hD0PNMu8CfQC.

Labov W (1986). ''The social stratification of (r) in New York City department stores.'' In ''Dialect and language variation,'' Elsevier.

Lambert WE, Hodgson RC, Gardner RC, Fillenbaum S (1960). ''Evaluational reactions to spoken languages.'' *The Journal of Abnormal and Social Psychology*.

Leerunyakul K, Suchonwanit P (2020). ''Asian hair: a review of structures, properties, and distinctive disorders.'' *Clinical, Cosmetic and Investigational Dermatology*.

Lev-Ari S, Keysar B (2010). ''Why don't we believe non-native speakers? The influence of accent on credibility.'' *Journal of Experimental Social Psychology*, **46**(6), 1093 -- 1096. ISSN 0022-1031. URL http://www.sciencedirect.com/science/article/pii/S0022103110001459.

Li J, Longinos G, Wilson S, Magdy W (2020). ''Emoji and Self-Identity in Twitter Bios.'' In ''Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science,'' URL https://www.aclweb.org/anthology/2020.nlpcss-1.22.

McGill A (2016). ''Why White People Don't Use White Emoji.'' URL https://www.theatlantic.com/politics/archive/2016/05/white-people-dont-use-white-emoji/481695/.

Miltner KM (2020). ''''One part politics, one part technology, one part history'': Racial representation in the Unicode 7.0 emoji set.'' *New Media & Society*.

Morioka T, Ellison NB, Brown M (2016). ''Identity Work on Social Media Sites: Disadvantaged Students' College Transition Processes.'' In ''Proceedings of

the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing,'' URL https://doi.org/10.1145/2818048.2819959.

Office for National Statistics (2011). *2011 UK Census aggregate data*. URL http://dx.doi.org/10.5257/census/aggregate-2011-1.

Ogbonnaya-Ogburu IF, Smith AD, To A, Toyama K (2020). ''Critical Race Theory for HCI.'' In ''Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,'' CHI '20, p. 1–16. Association for Computing Machinery, New York, NY, USA. ISBN 9781450367080. URL https://doi.org/10.1145/3313831.3376392.

Pardes A (2015). ''The Solution to the Emoji Diversity Problem: Make Them All Yellow.'' URL https://www.vice.com/en_uk/article/wd7ejm/emoji-shouldve-made-all-their-characters-yellow-408.

Pavalanathan U, Eisenstein J (2015). ''Audience-modulated variation in online social media.'' *American Speech*, **90**(2), 187--213.

Princewill V (2017). ''Is it OK to use black emojis and gifs?'' *BBC News*. URL http://www.bbc.co.uk/news/av/world-40931479/is-it-ok-to-use-black-emojis-and-gifs.

Quintana D, Williams D (2018). ''Bayesian alternatives for common null-hypothesis significance tests in psychiatry: a non-technical guide using JASP.'' *BMC Psychiatry*.

Rabagliati H, Robertson A (2017). ''How do children learn to avoid referential ambiguity? Insights from eye-tracking.'' *Journal of Memory and Language*, **94**, 15--27.

Ratcliff R, Rouder JN (1998). ''Modeling response times for two-choice decisions.'' *Psychological science*.

Robertson A, Magdy W, Goldwater S (2018). ''Self-Representation on Twitter Using Emoji Skin Color Modifiers.'' *Proceedings of the International AAAI Conference on Web and Social Media*. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/15055`.

Robertson A, Magdy W, Goldwater S (2020). ''Emoji Skin Tone Modifiers: Analyzing Variation in Usage on Social Media.'' *ACM Transactions on Social Computing*. URL `https://doi.org/10.1145/3377479`.

Robertson A, Magdy W, Goldwater S (2021a). ''Black or White but never neutral: How readers perceive identity from yellow or skin-toned emoji.'' *CSCW*.

Robertson A, Magdy W, Goldwater S (2021b). ''Identity Signals in Emoji Do not Influence Perception of Factual Truth on Twitter.'' In ''Proceedings of the Fourth International Workshop on Emoji Understanding and Applications in Social Media,'' .

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009). ''Bayesian t tests for accepting and rejecting the null hypothesis.'' *Psychonomic bulletin & review*.

Rudman LA, Goodwin SA (2004). ''Gender differences in automatic in-group bias: Why do women like women more than men like men?'' *JPSP*, **87**(4), 494.

Schanda J (2007). *Colorimetry: understanding the CIE system*. John Wiley & Sons.

Sharma G, Wu W, Dalal EN (2005). ''The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations.'' *Color Research & Application*, **30**(1), 21--30. doi:10.1002/col.20070. `https://onlinelibrary.wiley.com/doi/pdf/10.1002/col.20070`, URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/col.20070`.

Shoemark P, Kirby J, Goldwater S (2017a). ''Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data.'' In ''Proceedings of the Workshop on Stylistic Variation at EMNLP 2017,'' Copenhagen, Denmark.

Shoemark P, Sur D, Shrimpton L, Murray I, Goldwater S (2017b). ''Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media.'' In ''Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers,'' volume 1, pp. 1239--1248.

Simpson GB (1994). ''Context and the processing of ambiguous words.'' *Handbook of psycholinguistics*, **22**, 359--374.

Smyth R, Jacobs G, Rogers H (2003). ''Male voices and perceived sexual orientation: An experimental and theoretical approach.'' *Language in Society*, pp. 329--350.

Stanley DA, Sokol-Hessner P, Banaji MR, Phelps EA (2011). ''Implicit race attitudes predict trustworthiness judgments and economic trust decisions.'' *PNAS*.

Statistics South Africa (2011). ''Census 2011.'' URL `http://www.statssa.gov.za/?page_id=3836`.

Staum Casasanto L (2010). ''What do listeners know about sociolinguistic variation?'' *University of Pennsylvania Working Papers in Linguistics*.

Thelwall M, Buckley K, Paltoglou G (2012). ''Sentiment strength detection for the social web.'' *Journal of the Association for Information Science and Technology*, **63**(1), 163--173.

Tutt P (2015). ''Apple's new diverse emoji are even more problematic than before.'' *The Washington Post*. URL `https://www.washingtonpost.com/`

posteverything/wp/2015/04/10/how-apples-new-multicultural-emojis-are-more-racist-than-before/.

United States Census Bureau (2010). "Census 2010." URL https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045217.

Valla LG, Bossi F, Calì R, Fox V, Ali SI, Rivolta D (2018). "Not only whites: racial priming effect for black faces in black people." *Basic and Applied Social Psychology*.

van Doorn J, van den Bergh D, Böhm U, Dablander F, Derks K, Draws T, Etz A, Evans NJ, Gronau QF, Haaf JM, *et al.* (2020). "JASP guidelines for conducting and reporting a Bayesian analysis." *Psychonomic Bulletin & Review*.

Weidmann NB, Benitez-Baleato S, Hunziker P, Glatz E, Dimitropoulos X (2016). "Digital discrimination: Political bias in Internet service provision across ethnic groups." *Science*, **353**(6304), 1151--1155.

Weiß M, Bille D, Rodrigues J, Hewig J (2020). "Age-Related Differences in Emoji Evaluation." *Experimental Aging Research*, **46**(5), 416--432. doi:10.1080/0361073X.2020.1790087. PMID: 32662319, URL https://doi.org/10.1080/0361073X.2020.1790087.

Wood MA, Bukowski WM, Lis E (2016). "The digital self: how social media serves as a setting that shapes youth's emotional experiences." *Adolescent Research Review*.

Zimmerman J (2015). "Racially diverse emoji are a nice idea. But will anyone use them?" *The Guardian*. URL https://www.theguardian.com/commentisfree/2015/apr/22/racially-diverse-emoji-are-a-nice-idea-but-will-anyone-use-them.