SOCIAL MEDIA AND THE GENERATION, PROPAGATION, AND DEBUNKING OF RUMOURS

by:

Claire Baxter, Patricia Barratta, & Michael Thomson

Human*Systems*® Incorporated
111 Farquhar Street
Guelph, ON N1H 3N4

HSI® Project Manager:
Michael H. Thomson
(519) 836-5911 ext. 301

PWGSC Contract No. W7719-135235/001/TOR
Task Authorization No. 5235-08

On Behalf of

DEPARTMENT OF NATIONAL DEFENCE

As represented by

Defence Research and Development Canada Toronto Research Centre
1133 Sheppard Ave West
Toronto, ON M3K 2C9

Project Scientific Authority:Afzal Upal, Defence Scientist, 416-635-2170

March 2015

The scientific or technical validity of this Contract Report is entirely the responsibility of the Contractor and the contents do not necessarily have the approval or endorsement of the Department of National Defence of Canada

Contract Report
DRDC-RDDC-2015-C140
July 2015

Author

Michael Thomson
Human*Systems*® Incorporated

Approved by

Afzal Upal
Scientific Authority

Approved for release by

Name of Document Review Chair
Chair, Document Review and Library Committee

Abstract

The following work is in support of the Canadian Armed Forces C4ISR project “Socio-Cultural Intelligence Support to Joint Tactical Targeting.” The first aim was to understand the relationship between rumours and social media. The scientific literature was examined to identify the mechanisms of rumour generation, propagation and debunking. Based on the literature review, a framework was developed to examine particular case studies in the literature. A second aim of this work was to highlight effective techniques that outgroup members can use to influence core social identity beliefs of a target group. Ineffective techniques were also identified. Recommendations for future scientific work are included.

Executive Summary

SOCIAL MEDIA AND THE GENERATION, PROPAGATION, AND DEBUNKING OF RUMOURS

Claire Baxter, Patricia Barratta, & Michael Thomson, **Human*Systems*® Incorporated; DRDC Toronto CR2015-XXX; Defence R&D Canada – Toronto; March 2015.**

The following work is in support of the Canadian Armed Forces C4ISR project "Socio-Cultural Intelligence Support to Joint Tactical Targeting." The first aim of this work was to understand the relationship between rumours and social media. To this end, the scientific literature on rumours was examined to identify the mechanisms of rumour generation, propagation and debunking. Based on the literature review, a framework was developed to examine particular case studies in the literature.

The conditions that appear necessary for rumour generation include uncertainty, a threat to oneself or others, and the level of importance the situation/event is to oneself. Rumor generation is also motivated by sense-making, relationship enhancement, and self-enhancement. There are three primary types of rumours, including a wish rumour, dread rumour, and hostile rumour. The latter can be either intentional or unintentional. Propagation of rumours can occur more quickly and more broadly through social media. And they gain better traction if they are propagated by a credible source and are more believable. Confirmation bias and sense-making play a role in the propagation of rumours as well. Once people start to believe something, they are more inclined to search for evidence that confirms their beliefs (Nickerson, 1998). To debunk a rumour, the strategies to use include counter rumouring with clear evidence, logical refutation, and the addition of disconfirming information. Humour is also a possible way to expose a rumour and its false content.

There are five case studies that are examined, regarding the generation, propagation and debunking. These include rumours across social media during the 2011 UK riots, the Haiti earthquake, the Mumbai terrorist attacks, the Boston Marathon bombings, and the death of Islamic terrorist Noordin Mohammed Top. When possible, the framework is employed to highlight elements of rumour generation, propagation and debunking in the case studies. Recommendations for future scientific work are provided.

A second aim of this work was to highlight effective techniques that outgroup members can use to influence core social identity beliefs of a target group. To do this, communications need to overcome the intergroup sensitivity effect, which is the tendency of individuals to react more defensively to criticism made by an outgroup member compared to an ingroup member even when the message is identical (Hornsey & Esposo, 2009). Effective techniques to decrease defensive reactions to outgroups and increase attitude change include adding praise to criticism, acknowledge own group's shortcomings, attributing shortcomings to internal rather than external factors, and invoking a common identity (i.e., making identification with the target audience salient). Argument quality, identifying a subset of the ingroup, and being a powerful outgroup were not effective persuasive techniques.

Table of Contents

This page intentionally left blank.

1. Social Media and Rumours

The following chapter considers three core objectives. First, it considers rumour generation, specifically the techniques used to mobilize individuals to engage in the generation and propagation of favourable social media rumours and the discrediting of unfavourable social media rumours. Second, it considers case studies centring on the generation, propagation, and debunking of social media rumours. Finally, this report considers ways to support and grow Canadian Armed Forces' capacity to debunk rumours generated by enemy forces and ensure that these efforts reach the target audience.

1.1 Understanding Rumours

Everyone is familiar with rumours. These are "unverified information statements that circulate about topics that people perceive as important; arise in situations of ambiguity, threat, or potential threat; and are used by people attempting to make sense or to manage risk" (DiFonzo, 2008, p. 38). And Rumours can "impose real damage on individuals and institutions," if they are not either confirmed as true or discredited as false (Sunstein, 2009, p. 3). It is important, therefore, to understand how rumours are generated, how they gain traction, and how they are debunked.

1.1.1 Rumour Generation

It is argued that rumours are often generated (and propagated) in situations that are important, uncertain and threatening or uncontrollable and produce anxiety (Allport & Postman, 1947a; Rosnow, 2001, DiFonzo & Bordia, 2007; DiFonzo, 2008). For example, rumours may often be generated in wars or crises because they are life or death situations, and are certainly threatening, uncontrollable and anxiety-producing. Wars or crises are also uncertain situations. Accurate and complete information may be unavailable or late coming, which is the perfect condition for rumour generation. Although wars and crises are important, they are most important to those involved, suggesting the importance is relative. According to Kelley (2004), situations are important to an individual when the outcome of a situation is relevant to that person. As Allport and Postman (1947b, p. 502) stated, "an American citizen is not likely to spread rumors concerning the market price for camels in Afghanistan because the subject has no importance for him." Again, conditions that are probable for rumour generation (and propagation) are importance, uncertainty, threatening, uncontrollable, and anxiety producing.

Within these conditions, people are motivated to make sense of the situation (DiFonzo, 2008). For example, in the immediate aftermath of a terrorist attack, people have many questions that are unanswered. Why was there a terrorist attack? What happened? Who is responsible? Will there be more? Where can one get help? To make sense of these events a rumour may be generated (and propagated). It is likely that the generation of these rumours are, more often, unintentional and the result of collective discussions around reasonable and hedged guesses to fill in the blanks (Kapferer, 1992). According to DiFonzo and Bordia (2007, p. 72) "the informal interpretation arising out of this collective process becomes a rumor." This sense making motivation is highest after an event has occurred, with decreased interest as time goes on (Shibutani, 1966). So sense making is a particular motive to generate and spread rumours.

But people also generate rumours to fulfill relationship-enhancing motives as well as self-enhancement motives (DiFonzo and Bordia, 2007). With respect to the former, providing information to others, rumour or not, is a way of increasing one's status in a group (Brock, 1968; Fromkin, 1972;

Lynn, 1991). People may unknowingly pass on incorrect information to a group to enhance their social status in that group. And research shows people are more likely to tell another person a positive rumour than a negative rumor to invoke positive affect in another individual (Kamins, Folkes, & Perner, 1997).

Rumour generation can also serve self-enhancement motives, i.e., to feel better about oneself, and these can be unintentional and intentional (DiFonzo & Bordia, 2007). Unintentional rumours spread to fulfill self-serving motivations can be in the form of positive ingroup rumours or negative outgroup rumours. For example, we can increase our self-esteem by generating and propagating positive ingroup rumours and negative outgroup rumours. As DiFonzo and Bordia (2007, p. 79) explain, "spreading rumors may boost one's self-esteem by boosting one's social identity."

If the outgroup is one which is the target of prejudice , then the rumours may also be spread to "rationalize self-enhancing attitudes" (DiFonzo & Bordia, 2007, p. 80). Rumours that justify existing prejudices are more likely to be spread than those that are counter to our prejudices. According to Kunda (1990, p. 483), rumours serve in the process of "justification construction." While justifying prejudices, the generation and propagation of these types of self-enhancing rumors may not be entirely intentional.

Intentional self-enhancing rumours, on the other hand, have been spread in situations such as war in the form of propaganda (Allport & Postman, 1947b; Rosnow, 2001) or to demoralize enemy troops (Allport & Postman, 1947b; Mihanovic, Jukic, & Milas, 1994) or as part of a tactile maneuver (e.g., Nazi Joseph Goebbels spread rumors about German operations to function as a smoke screen) (Allport & Postman, 1947b; DiFonzo, 2008). Propaganda rumours are "misinformation deliberately planted to gain political, strategic, competitive, or military advantage" (DiFonzo, 2008, p. 23). They can occur in elections (Kapferer, 1990), be part of sales tactics (Kapferer, 1990; Turner, 1993), or even to bring down stock sales (Kapferer, 1990). Self-enhancement rumours are "the least conducive to accuracy, yet often have the most damaging effects such as instigating hostile reactions or inter-group conflict" (Kelley, 2004, p. 19). This is primarily because people are unwilling to generate or propagate rumours that violate their worldview that their ingroup is better than an outgroup (Kelley, 2004).

Some individuals seem to delight in the production and dissemination of intentionally malicious rumours. Indeed, "trolls" or "those who provoke other [internet] users and disrupt discussion; posting off-topic or making inflammatory statements" have been identified as a "considerable source of annoyance" (Taylor, 2012, p. 25). This was particularly true for a team who developed a Facebook site for the dissemination of useful information and support for victims of Cyclone Yasi in Australia. The trolls seemed to "delight in crushing goodwill and attempts to help others" (Taylor, 2012, p. 25).

Now that we have identified the conditions and the motives of rumours, it is necessary to identify the different types of rumours. The most common are wish, fear/dread, or hostility/wedge-driving rumors. According to DiFonzo (2008), a wish rumour centres on a hoped-for outcome. For example, just before the end of WWII, DiFonzo (2008) reports that many wish rumours were circulating that the war had ended. A dread rumour is created under similar conditions but concerns a feared negative outcome (DiFonzo, 2008). During the SARS outbreak there was an information gap that was filled with rumours about potential severity and spread of SARS (DiFonzo, 2008). The third type of rumour, hostile or wedge-driving rumours, "feed on hate and serve to drive a wedge between people by derogating another group" (DiFonzo, 2008, p. 17).

Today, the form of rumours has been impacted by the Internet and social media. In a move from a read-only Web to a participatory read-write Web 2.0 (O'Reilly, 2007), whereby users both produce and consume media (prosumption; Bernardi, Cheong, Lundry, & Ruston, 2012), online rumours are not just in the form of a text or verbal rumour. They are accompanied by manipulated (or photoshopped) images and videos with the "intent to both shock and to persuade" (Bernardi et al., 2012, p. 49). For example, during Hurricane Sandy images were spread through social media, including sharks swimming in a flooded mall and a dramatic hurricane cloud over the Statue of Liberty (Gupta, Lamba, Kumaraguru, & Joshi, 2013). Thus, rumours are not merely in the form of words, but may also include manipulated images.

Now that we have looked at the generation of a rumour, it is necessary to look at the process of propagation.

1.1.2 Rumour Propagation

When rumours are propagated they do not always stay the same. They change or mutate. According to DiFonzo and Bordia (2007) rumours can change through leveling (loss of detail), sharpening (accenting or highlight parts of the rumour), adding (additional of content) and assimilation (changing rumour to fit in line with personal schemas). In real life situations, when given the freedom to discuss a rumour, DiFonzo argues that rumour distortion decreases and adding increases because "discussion involves repetition, rehearsal, meaningful encoding, active listening, message checking, and interaction; all of these features aid memory and boost attention" (DiFonzo, 2008, p. 161). In everyday life, when discussion is not restricted, rumours are most likely to sharpen and assimilate (Buckner, 1965; Peterson & Gist, 1951; Rosnow, 1991; Shibutani, 1966; Turner, 1964; 1994; Turner & Killian, 1972; all cited in DiFonzo & Bordia, 2007). People are more likely to accent or highlight aspects of a rumour to fit their personal schemas (a way in which people perceive a concept based on their knowledge and personal experiences; Fiske & Taylor, 1991).

When Allport and Postman (1947a) considered rumours, more than a half century ago, they were thought to be spread primarily via word of mouth. Today, the transmission of rumours extends to print (e.g., newspapers), electronic media and the Internet (Rosnow, 2001). In fact, with the move from a read-only Web to a participatory read-write Web 2.0 and the advent of social media, the rumour has been significantly impacted in terms of its propagation.

The Internet, or Web 2.0, where the traditional dichotomy of media producers and media consumers has broken down has seriously impacted the propagation of rumours. Messages spread online are transmediated, meaning these messages are appropriated (people can copy and paste text, download photos and videos related to rumours and claim them to be their own), reconfigured (text, images and videos can all be changed in terms of leveling, sharpening, adding, and assimilating) and retransmitted across different media platforms (Bernardi et al., 2012). And according to Bernardi et al. (2012, p. 171), "each change of medium involves both alteration of form and, sometimes subtly, meaning."

Not only does the online environment impact rumour propagation, so do the variety of interfaces of social media[1] (Friggeri, Adamic, Eckles, & Cheng , 2014). However, the ways in which they change depend on the interface of the social media program. Unlike rumours spread by word of mouth or text, where there are few limits on what or how much is said, the different social media interfaces vary in the extent to which users are freely able to change the original message that they are reposting, sharing or retweeting.

[1] Social media are "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 that allow the creation and exchange of user-generated content" (Kaplan & Haenlein, 2010, p. 61). Currently, the most popular social media websites are, arguably, Facebook, Twitter, Instagram, Pinterest, Tumblr, Google+, YouTube and Reddit. Facebook, Twitter and Instagram are typical of social networking sites, where users create their own online social network by linking to their friends' accounts or following other users. Users post messages or images that are broadcasted to their social network. Unlike Facebook and Google+, Twitter, Instagram, Pinterest, YouTube and Tumblr have a directed and non-reciprocal friendship model (Procter, Vis, & Voss, 2013), such that users can follow any other user that they choose, but these users do not have to follow them back. In other words, a user cannot control who follows them.

Over the years, Facebook has changed the ways users can spread information to their social network or public. A few years ago, when users wanted to spread a post written by their Facebook friend, they had to copy and paste it and post it on their Facebook page. Recently, Facebook has introduced a "Share" button so that a user can easily share their friend's post without copying and pasting. The share function automatically copies and pastes their friend's post to their page. A consequence of this change is that users cannot change their friend's original post, similar to serial transmission. However, they can add a comment alongside the post. To edit the original post, one must use the traditional copy and paste method, which is still available.

On Twitter, a user can retweet another user's tweet. The original tweet appears on the second user's page verbatim. This is the typical sharing method on Twitter. However, there are also ways to retweet with a comment, if sharing with a specific other with a direct (and private) message. Like Facebook, the option to paraphrase or copy and paste is always available and as a result, the message can change. Because Twitter and Facebook are used frequently as an app on phones, the copy and paste method may be used less frequently as this function is not as user friendly as it is on a computer. It is possible that the share function on Facebook may impact the extent to which rumours mutate.

Propagation is also impacted by the source characteristics. For example, recipients are more likely to attend to messages from a credible source and dismiss messages from a noncredible source (Kumkale & Albarracin, 2004; Petty & Cacioppo, 1986). Misinformation that is disseminated from a credible source is more likely to be propagated as a rumour than from a noncredible source (DiFonzo & Bordia, 2007; Caplow, 1947; Guerin, 2003, Stevens & Fiske, 1995; all cited in DiFonzo & Bordia, 2007). According to McCroskey and Teven (1999), source credibility includes an assessment of the source's competence, goodwill, and trustworthiness. In a conflict, of course, trust is often hard to come by (Kelley, 2004). Often during conflict, an authority is not removed from the rumour. It is perceived to be impacted by the public's belief or disbelief in the rumour. Further, the message may be coming from an authority or person be perceived as part of an outgroup, particularly in divided countries or regions. Messages from an outgroup are perceived as less credible (Petty & Cacioppo, 1986).

The pseudonymous environment of social media can impact the availability of the source and, consequently, the source's credibility (Ozturk, Li, & Sakamoto, 2015). Within social media, even though the person may use their real name, the original source of the message is more likely to be obscured. Although many rumours can be retweeted or shared from the original source, it only takes one user to use a copy and paste method or to paraphrase the original rumour for the source of the post to be removed. When the message is stripped from its original source, it can increase the credibility of a source with low credibility or has the potential to suppress skepticism that would normally accompany messages from the 'enemy' (Kelley, 2004). Furthermore, when a message has been retweeted or shared, "it is tacitly being endorsed by the forwarder. And when it has been forwarded by so many people, it may gain a certain degree of credibility, and in this way increase the likelihood that it will be forwarded yet again" (DiFonzo, 2008, p. 107). In other words, "it is possible for malicious users to leverage the trust we have in our 'friends' or 'connections' on social networking sites in order to spread harmful content" (Tripathy, Bagchi, & Mehta, 2013).

Rumour characteristics also have an impact on their propagation. For example, rumours that are believable are more likely to be spread than less believable rumours (DiFonzo, 2008). Further, the more times a rumour is heard, the more believable it is. Rumours "are unlikely to take hold unless they are perceived as plausible, and the more a specific rumor is heard the more currency it gains, shifting the orientation of public thinking and ultimately steering opinion" (Kelley, 2004, p. 6). The more fit a rumour or a piece of cultural information is, the more it propagates (Dawkins, 1976).

Rumours that are generated for particular motivations (e.g., sensing making, relationship-enhancing, or self-enhancing) are also propagated because they serve those motives (DiFonzo & Bordia, 2007). Although this distinction is not entirely clear in the literature, we believe that it is likely that rumours that serve the original motives in rumour generation will also be propagated more. People are more likely to listen and be influenced by messages that are communicated in a way which matches their motivation (Cesario, Grant, Higgins, 2004).

According to DiFonzo (2008), network structure, or social topography, influences propagation of rumours too. People can be in groups of a highly segregated nature (e.g., family units) or non-segregated/random nature (connections cannot be predicted, such as a group of strangers). The more segmented a population, the more diversity there is in the rumour because they are shielded from exposure to "global majorities – and the social pressures that go along with this exposure" (DiFonzo, 2008, p. 119). On the flipside, the less segmented a population, the less diversity there is in a rumour. Social media may conform more to a segregated network if we tend to follow or friend those who are more alike. On the other hand, it may be that we have greater access to diversity on social media because social networks enable many-to-many conversations (Joseph, 2012).

1.1.3 Rumour Debunking

Once a rumour is established and propagated, how does it get debunked? A rumour helps a person make sense of an event. Counter information may have a diluted impact because the person feels they already understand the situation (Johnson & Seifert, 1994). Referred to as the confirmation bias, people have a tendency to try to confirm their beliefs rather than disconfirm their beliefs (Nickerson, 1998). If people believe they understand the situation, therefore, they may be more inclined to pursue information, interpret information, or recall information that confirms this belief (Nickerson, 1998). When people are building mental models of information, disconfirming information is more likely to have an impact when the model is being built rather than after it is solidified (Johnson & Seifert, 1994). When counter information is late there is a continued influenced effect, which is the persistent reliance on this misinformation, even when later people are presented with a correction or retraction (Johnson & Seifert, 1994).

Research has demonstrated that, to some degree, rumour rebuttals decrease belief in rumours (Allport & Postman, 1947a; DiFonzo & Bordia, 2004), whereas other studies suggest otherwise (Ozturk & Sakamoto, 2015). Even in real-life situations, direct rebuttals have not been met with much success. During WWII, the U.S. Government developed a radio broadcast to debunk wartime rumours. It showed that most listeners tuned in to listen to the rumour and then changed the station before the rebuttal was provided. Additionally, Allport and Postman set up rumour clinics in WWII with the purpose of debunking rumours. Instead of debunking rumours with direct information (which may have been classified), Allport and Postman tried to explain the psychological underpinning of belief in rumours (e.g., self-defense, mental projection) and were met with little success.

DiFonzo (2008) explains that his approach to countering rumours is motivated by a desire to persuade individuals with clear and strong evidence as the process unfolds in order to help them make good sense of the situation. However, strong evidence is sometimes difficult for an official source to produce. Information may be classified or may interfere with interests of national security (Bernardi et al., 2012). Further, the disconfirming evidence may shift and change depending upon organizational/governmental intentions (Benardi et al., 2012). For example, coalition forces in Iraq found it difficult attempting to deny rumours related to reasons why they were in Iraq, because the official reasons changed from "weapons of mass destruction to establishing freedom and democracy in the Middle East to national building" (Bernardi et al., 2012, p. 94).

The information that the rumour is countered with can comprise facts, arguments, and psychological mechanisms. For example, the rumour that President Obama was not born in the U.S.A. can be countered with direct information in the form of fact: his birth certificate. It could have also been countered with an argument such as Obama's mother and grandparents (who raised him) lived in America, so it is likely that Obama was born in America. It could have also been countered with explaining the psychological mechanism underlying the propagation of the lie, such as motivations (e.g., Donald Trump is trying to propagate this rumour because he wants a Republican in power and he himself wants to be President) or prejudice (e.g., people spreading this rumour cannot accept a Black man in power).

People can also counter a rumour by suggesting that they heard the counter information from a credible source. However, when people do not provide proof of their source (fail to provide a link) they are essentially asking people to "take my word for it." People do also counter rumours in a "take my word for it" approach by just pointing out the rumour is false without providing an explanation of why.

Another means of debunking a rumour is through a counter rumour. Tripathy and colleagues (2013) argue that rumours are best countered, not with direct (and boring) countering information, but with an anti-rumour. In other words, "in situations where populations do not answer to the same authority, it is the trust that individuals place in their friends that must be leveraged to fight rumor. In other words, a rumor is best combated by something which act [sic] like itself, a message which spreads from one individual to another" (Tripathy et al., 2013, p. 149).

Friggeri and colleagues (2014) investigated the propagation of user generated counter rumours within Facebook. They found that a counter rumour needs to be just as titillating as the rumour it is countering because countering a rumour with facts is dull. For example, in response to the rumour that Facebook would become a pay site, a general statement against rumour propagation was generated. It said, "Don't blindly copy and paste warnings just because your Facebook friend's status tells you to do so. Although you probably mean well, you could be helping a hoax become more popular" (Friggeri et al., 2014, p. 8). This post only generated 5000 text copy and pastes (reposts). However, a more titillating counter rumour went "viral". It said the following (Friggeri et al., 2014, p. 9):

> "On September 31st 2011 Facebook will start charging you for your account. To avoid this you must get naked, stand on your dining room table and do the Macarena all the while singing 'I will survive' after filming and posting it to your Facebook wall and YouTube then and only then will Mark Zuckerberg come down your chimney to tell you that your account will stay free. Pass it on it must be true because someone on Facebook I hardly know told me."

This counter rumour received wide distribution and along with two similar others, achieved over 8.6 million posts (similar to the 12.7 million posts of the original rumour; Friggeri et al., 2014).

It seems then that when countering a rumour it is important that one not only has an audience, but captivates the audience, not only by the topic, but by the form of the presentation of the countering message. As evident in the above example of the counter rumour, humour can help to keep people's attention.

Humour can be used to demonstrate the ridiculousness of the rumour itself. Bernardi and colleagues (2012) provided the example of an SNL skit used to indirectly debunk the misinformation spread by Iran's President Mahmoud Ahmadinejad who declared that there were no homosexuals in Iran. In the SNL skit, Andy Samberg romantically serenades Fred Armisen, who is impersonating Ahmadinejad. Additionally, ridicule can also be used as a way to reduce the credibility of the source of a rumour and indirectly help to counter the rumour (Bernardi et al., 2012). However, not all countering strategies can include the use of humour, particularly in sensitive situations. But still, captivating, or at least engaging an audience should be considered when countering rumours.

In the rumour literature, there are three models of rumour countering that have been statistically modelled (Tripathy et al., 2013): the delayed start model, the beacon model and the neighbourhood model. The first two models involve rumour countering from an official source, the third is a model involving rumour countering from the public. The delayed start model is a reactive model and suggests that a local authority discovers a rumour *n* days after it starts and begins its rumour countering strategy relatively late. According to Tripathy et al. (2013, p. 151), the beacon model is a proactive model and "models a situation where a set of vigilant agents, *beacons*, are on the lookout for the spread of rumours." Immediately after rumour identification, rumour countering occurs (Tripathy et al., 2013). However, a statistical analysis found that a neighbourhood model (Tripathy et al., 2013) or a self-correcting crowd (Mendoza, Poblete, & Castillo, 2010) is as effective (if not more) than either the delayed start or the beacon model (Tripathy et al., 2013). A neighbourhood model is a "non-authority centric model of an enlightened citizenry vigilant against rumours, in which any use with some probability can detect the rumor on receiving it and decide to warn his or her contacts about the spread of the rumor" (Tripathy et al., 2013, p. 151).

According to DiFonzo (2008), rumour debunking occurs most effectively when it is refuted early by a trusted source. But officials experience a barrier due to potential distrust from the public. Trust is difficult to come by in conflict (DiFonzo, 2008). An official source may be perceived as having vested interest (Tripathy et al., 2013) and may be perceived as an outgroup. As Kelley (2004, p. 53) suggests, "it is more effective to refute the rumour, and most effective to refute the rumour with a trusted source that does not have a vested interest in the subject." However, what is a trusted source may vary from person to person. Some may trust information from an official source, however, some may trust information more from their social network than from an official source because they may not have vested interest and are a part of one's ingroup.

Like official sources, the public or people in one's social network can debunk a rumour with direct information, by explaining the psychological mechanism (e.g., bias) or by propagating a counter rumour. People in one's social network can refute a rumour early, are trusted and may be perceived as an unvested source. This is likely why a neighbourhood model of counter rumour propagation is effective, if not more than the delayed start and beacon models (Tripathy et al., 2013). Based on these results, Tripathy et al. (2013) suggest that social networking users should be educated about the dangers of rumours and should be encouraged to prevent their propagation.

Researchers employed by Facebook found that Facebook users attempt to debunk their friends' rumours by posting links to Snopes.com, an internet site that documents the spread of rumours and guesstimates its truth value (Friggeri et al., 2014). Friggeri and colleagues (2014) found that false rumour posts were more likely to have a link to Snopes.com in the comments section than true rumours, and they are more likely to have such links shortly after being posted. Consequently, posts with false rumours (reshares) and links to Snopes.com in the comments were 4.4 times more likely to be deleted than when no link was provided. The researchers note that, even though these posts may have been deleted, the rumour continued to propagate online because non-snoped posts still existed. Even though a post was more likely to be deleted after being "snoped", most reshares occurred after the post was snoped in the comments. Friggeri et al. (2014) suggest that this might be because the snopes comment was hidden by the time the readers read the post or that they ignored the snope. "[L]arge cascades are able to accumulate hundreds of Snopes comments while continuing to propagate" (Friggeri et al., 2014, abstract). They also found that "rumour cascades run deeper in the social network than reshare cascades in general."

To debunk the rumour, we suggest early identification and refutation in accordance with the beacon model, from high source credibility, with clear and strong evidence, and a captivating presentation.

1.1.4 Rumour Framework

Based on the academic literature explored above on the generation, propagation, and debunking of rumours, we developed a framework to guide our investigation of case studies involving rumours and social media.

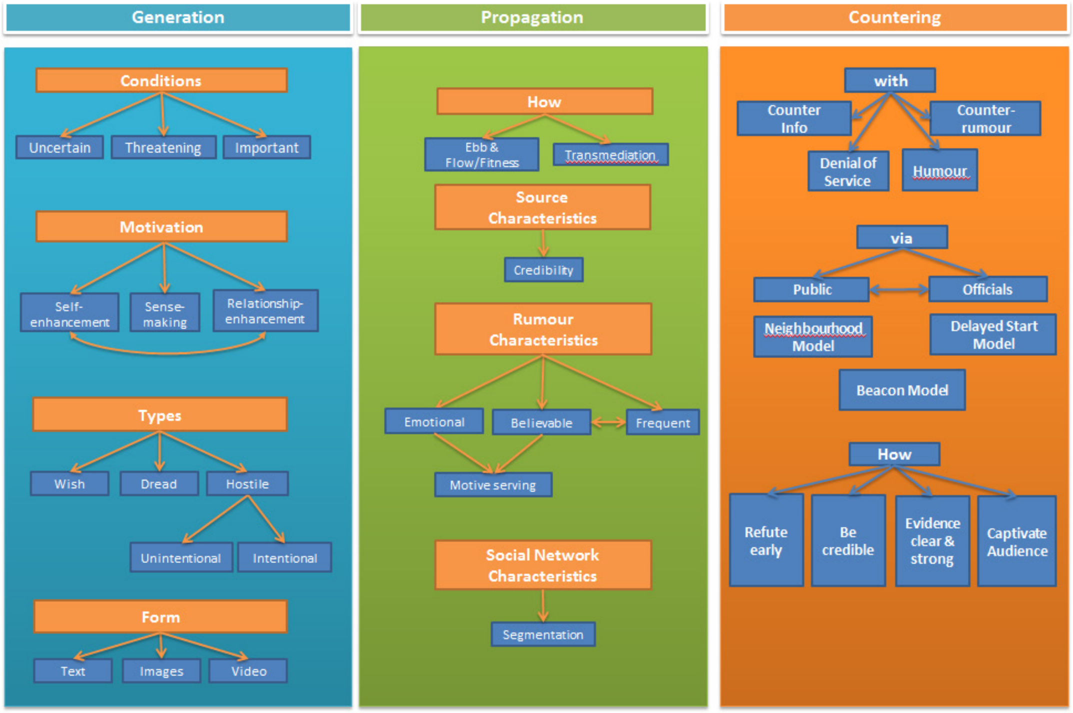

Figure 1: Generation, propagation and countering of rumours framework

1.2 Case Studies

Five case studies are presented here to demonstrate how rumours generate, propagate and are countered in social media in real life scenarios. The case studies include the UK riots; the natural disaster in Haiti; the Mumbai terrorist attack; the Boston Marathon bombings; and the discrediting of terrorist, Noordin Mohammed Top. Case study 1 of the UK riots focuses on rumour countering and the beacon model. Case study 2 of the Haitian earthquake focuses on rumour countering. Case study 3 of the Mumbai terrorist attacks in 2008 and 2011 focuses on propagation and rumour characteristics. Case study 4 of the Boston Bombing focuses on rumour propagation and transmediation. Case study 5 of the capture and execution of terrorist Noordin Mohammed Top focuses on counter rumour generation by an authority.

1.2.1 The UK Riots 2011

On the 4th of August 2011, notorious gang member, Mark Duggan, was shot dead by police in his home town of Tottenham, North London, England by the Metropolitan Police (Panagiotopoulos, Bigdeli, & Sams, 2014). Based on uncertainty over whether or not Duggan was armed and concerns of police brutality, a peaceful protest occurred two days later (Panagiotopoulos et al., 2014). However, after dissatisfaction with the response from senior police officers, thousands of people broke out in a riot in Tottenham (Panagiotopoulos et al., 2014). Over the course of six days, the riots spread across several London boroughs and some towns throughout England, such as Nottingham and Manchester. In the aftermath of the riot, five people had been killed ("Getting to the root," 2011) and 3,051 people faced riot-related charges (Bowcott, 2012).

Once the riots commenced and started to spread outside of London, rumours were spread about where these new riots were occurring (Denef, Bayerl, & Kaptein, 2013, Panagiotopoulos, Bigdeli, & Sams, 2012, Panagiotopoulos et al., 2014, Procter, Crump, Karstedt, Voss & Cantijoch, 2013, Tonkin, Pfeiffer, & Tourte, 2012; Vis, 2012). Other more sensational rumours were that riots were occurring at the Birmingham Children's Hospital and that the London Eye was on fire (Procter, Vis, & Voss, 2013). Research shows how social media was used to counter or debunk the rumours by appealing to both logical argumentation and counter information as well as a "take my word for it" approach, and in some cases humour.

For example, Procter, Vis and Voss (2013, p. 208) share a tweet from the riots that appeals to logical arguments to counter the rumour that the London Eye was on fire. It reads as follows:

> The London Eye isn't on fire, neither is Big Ben. You can't burn metal and heard of Photoshop, hellooooo? #londonriots

The rumour that the London Eye was on fire was primarily countered with attempts at logical arguments that the metal would need a much higher temperature to catch fire (Guardian Interactive Team, Proctor, Vis, & Voss, 2011) or that "you can't burn metal" (Procter, Vis, & Voss, 2013, p. 208). The rumour eventually died out on Twitter that day.

Another tweet from the Procter, Vis, and Voss research includes new information to counter the rumour that rioting was occurring at the Birmingham Children's Hospital.

> #birminghamriots brmb ratdio and chief medical officer have confirmed Birmingham children's hospital has NOT been hit by riots

A local authority uses more of a "take my word for it" approach.

According to Guardian Interactive Team et al. (2011) the rumour that rioters were attacking the Children's hospital in Birmingham was countered by users within the first few hours of its propagation. One user offered counter information about the rioting at the Children's Hospital, and tweeted that his girlfriend worked there and has not seen any rioting, thereby invoking a level of credibility. Another user appealed to logic stating that the Children's hospital is across the street from a police station and is thus an unlikely venue for a riot. Another user invoked authority, stating that the Chief Medical Officer denied the rioting claim. The rumour petered out the next day (Guardian Interactive Team et al., 2011).

Tweets about rumours actually emerged across the country to diminish general fears and concerns. For example, the tweets below show individuals adopting a "take my word for it" approach.

> Load of rumours spreading about disorder across Sussex but all unfounded: business as usual for residents and visitor (Panagiotopoulos et al., 2014, p. 354).
>
> rumours suggesting disorder in Wycombe are simply not true #londonriots #carcrime

A study was conducted in a joint effort with journalists, designers and developers from the Guardian and researchers Procter, Vis and Voss that assessed how the more salacious UK riot rumours were countered by the public on Twitter (see http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter for an interactive display demonstrating concurrent rumour spread and rumour countering on Twitter) (Guardian Interactive Team, et al., 2011).

According to Guardian Interactive Team et al. (2011), a rumour that rioters had released a tiger from the zoo was defused(?) by a user with counter information. He suggested that the photos that accompanied the rumour on Twitter were actually from an incident in Italy where a tiger escaped and took to the streets. This counter argument was propagated by other users. It gained momentum when users used the rumour for humour purposes, reporting that they hoped the tiger would eat the rioters. Facts and humour managed to quell the rumour. The rumour eventually died out on Twitter the next day (Guardian Interactive Team et al., 2011).

According to the Guardian Interactive Team (2011), a rumour that rioters were cooking their own food in McDonalds was not countered on Twitter during their sampling timeframe. Similar to the tiger rumour, the rumour continued to propagate in the form of humour as one user was widely retweeted by saying that if she were to make her own food, she wouldn't have broken in to McDonalds (Guardian Interactive Team, 2011). The rumour died out on Twitter the next morning.

Counter information and logical argument were used to debunk a rumour that tanks were deployed at a bank in London. In fact, this rumour began with a number of queries (e.g., "is this real") from its start, suggesting that this was because the rumour was more unbelievable than others (Guardian Interactive Team et al., 2011). Using counter information, one Twitter user claimed the picture was actually of a tank in Egypt, whereas another user employed logical argumentation claiming it would not be a tactical advantage to deploy forces at a bank. The rumour eventually died out on Twitter that day.

According to Guardian Interactive Team et al. (2011), the rumour that the riots started because the police had attacked a 16-year-old girl was an unsubstantiated rumour. The rumour spread for three days on Twitter with little countering from the public or authorities. Users eventually provided links to a video of a woman yelling at police that the victim is just a girl, but the girl or police cannot be seen in the video. The rumour died out on Twitter without much countering after three days (Guardian Interactive Team et al., 2011). It was never substantiated.

In an analysis of rumour counters from local government authorities during the riots, which cover a wide range of services, Panagiotopoulos et al. (2014) found that 8% of the tweets from local authorities during the UK riots were aimed at preventing rumours. For example,

> **User:** Kate Stalker @Stalksie **Date:** 10 Aug 2011 **Time:** 12:16PM
>
> @nottspolice Is the rumour about Bramcote Lane shops being attacked tonight true? #Nottingham #riots #nottspolicelatest
>
> **User:** Notts Police @nottspolice **Date:** 10 Aug 2011 **Time:** 12:52PM
>
> @Stalksie We have no reported incidents in Bramcote Lane tonight #Nottingham #riots #nottspolicelatest

Not only is this tweet an indication of the timeliness that the Nottinghamshire Police force took in responding to rumours, it also represents a deviation from the typical broadcast form of interaction in Twitter (Denef et al. 2013). Denef et al. argues that many police forces used Twitter as a forum to broadcast a message to the masses. For example, Panagiotopoulos et al. (2014) found that a local authority was countering rumours with a press release and advertised the press release on Twitter.

> Harlow Council leader statement on riot rumours http://t.co/ilDvDWf.

Panagiotopoulos et al. (2014) found that over 25% of all tweets from local government authorities were direct replies to other users, typically people requesting information and that “most replies were related to rumors and even included warnings to specific users against the spreading of rumors” (p. 353). One tweet provided by Panagiotopoulos et al. (2014, p. 353) demonstrates how a local authority countered a rumour by directly tweeting the user who (likely) propagated the rumour. In the same tweet, they implored others to resist spreading rumours:

> @ [The User] no disorder in #Rugbytown at all. All quiet last night and tonight. Pls don't repeat unsubstantiated rumours!

It seems this type of encouragement to resist spreading rumours was typical. In an analysis of the top 25 retweeted messages, Procter and colleagues (2013) found that two tweets (similar in message) from centrally managed police force accounts countered rumours with direct information and appealed to the public to not listen to rumours and asked them to retweet their message. For example,

> Pls don’t listen to rumours! No #riots /damage in #Leicester we’ve placed additional staff as precaution only. Will keep you updated PRT

Another example of responding directly to the public is a tweet from a central police force Twitter account, likely either because they had spread a rumour or because they had requested information.

> @reeshasykes no incidents of disorder at #highcross #Leicester #riots

Not only did the police directly address individuals who were spreading rumours and respond to requests for clarification, they also requested Twitter users to update them on new rumours. For example, a Twitter account from a local police force tweeted the following:

> Any rumours of disorder in the South West no matter how believable please let us know asap 09452 777444 #ukriots #Devon #Cornwall #police

Some police forces were taking a pro-active stance, establishing beacons to identify and counter riot-location rumours. Police were communicating with the public through their twitter account and they were also checking their mentions and @replies to check for the public reporting rumours to them. The public were reading the police Twitter page for accurate information and were also retweeting this information to their followers and sometimes redirecting their followers to follow local police accounts for accurate information as in the following tweet from Procter, et al. (2013).

> There are no #leedsriots or #manchesterriots. Follow @gmpolice and @PoliceEastLeeds for correct info on events.

However, Procter et al. (2013) showed that there was a lack of communication between groups created for the purpose of collecting and disseminating information during a crisis (e.g., @IDrioters, created to confidentially identify rioters, @ManchesterRiots, created to disseminate situation information). Procter et al. (2013) even showed that @IDRioters made an effort to reach police by mentioning them in their tweets, but the police did not respond back. Groups such as these are important because they have the high potential for rumour generation and propagation, and they are an appropriate targeted group for a countering message.

Again, it is difficult to address the efficacy of these different forms of communication, but Denef et al. (2013) suggest that there are both pros and cons of formal and informal communication on Twitter. On the one hand, Denef et al. (2013) suggest that an informal style on Twitter may lead to an increased following. Indeed, they found that the Greater Manchester Police, who adopted an informal style during the riots, increased their number of followers from approximately 23,000 to more than 100,000, making them one of the most followed police forces on Twitter (only second to the FBI). However, an informal approach also may result in an overstepping of boundaries and offending the public (Denef et al., 2013). Procter et al. (2013) argue that police forces need a social media strategy that instead of using it as a source for broadcasting, use it as a capacity to inform about risk. Procter et al. (2013) explain that the Twitter accounts created at the time of the riots for the purpose of disseminating riot information sought engagement with police. As such, this could have been further exploited as a way to crowd-source situation information and debunk rumours.

1.2.2 The Haiti Earthquake

After the 7.0 earthquake hit Haiti on January 12, 2010, approximately 160,000 people died and approximately a quarter of all residents in Haiti had their homes completely destroyed (Kolbe, Hutson, Shannon, Trzcinski, Miles, Levtiz, et al., 2010). During this time, false rumours were spread that UPS would ship any package to Haiti if it was under 50 pounds and that a number of airlines would transport medical personnel to Haiti for free (Oh, Kwon, & Rao, 2010; Leberecht, 2010). The latter rumour also included the telephone number of the Haitian Consulate in New York which resulted in tying up the consulate phone line (Heussner, 2010). According to Oh et al. (2010), these rumours “turned out to be hearsay rather than eyewitness accounts, and subsequently refuted by UPS and airline companies as false information” (p. 2).

Oh et al. (2010) examined the propagation of rumours related to Haiti earthquake on Twitter through a sample of tweets that included the hashtag #haitiearthquake. They coded all English tweets in their sample related to the Haiti earthquake in terms of authenticating statements (where the user attempted to add credibility to their statement by citing a credible source or proclaiming to be an expert), emotional statements (including both positive and negative feelings), and work statements (asking others to do something like donate to charity), among others. They found that as authenticating statements increased on Twitter, the number of emotional statements decreased, suggesting that “the high levels of anxiety can be controlled at the early stage through feeds of governmental organizations, RSS, streaming videos, photo, text message, or Retweet, etc.” (Oh et al., 2010, p. 12). This suggests that reliable information from a credible source is a good way to manage the conditions essential for rumour generation, i.e., uncertainty, threat, etc. However, Oh et al. (2010) did not assess the impact of anxiety reduction on the propagation of a specific rumour. They found that an increase of authenticating statements seemed to occur at the same time that anxious statements decreased. Thus, Oh et al. (2010) extrapolated that since anxiety is associated with rumour generation and propagation (DiFonzo & Bordia, 2007), then a decrease in anxiety may lead to a decrease in rumours. Their data cannot fully support this conclusion.

Because there was limited amount of data on the effectiveness of the rumour countering strategy in Haiti, we decided to do a small scale investigation into how official sources countered rumours on Twitter. In our analysis of tweets from American Airlines between January 1, 2010 and March 1, 2010 that contained the word Haiti, we found that American Airlines countered the following medical personnel tweet (twitter.com, 2015):

> American Airlines @AmericanAir 13 Jan 2010 7:58 PM

> PLS RD: Cannot fly individual drs/nurses to #haiti, working w/Red Cross & other agencies 2 provide aid. Donate http://bit.ly/4zOgi0

This tweet was retweeted 285 times. However, we noticed that they used the hashtag #Haiti, which Oh et al. (2010) noted was filled with too many irrelevant tweets. This AA tweet did not quell all rumours, as 44 minutes later, @TheLoneOlive (with currently 3917 followers) tweeted the following (twitter.com, 2015):

> Amanda Lin Costa @TheLoneOlive 13 Jan 2010 8:44 PM
>
> confirmed American Airline flying relief missions as well as offering frequent flyer miles for donations:: #haiti

No less than one minute later (although it looks like one minute beforehand, but this is likely due to inaccuracies in Twitter's record keeping) American Airlines responded with:

> American Airlines @AmericanAir 13 Jan 2010 8:43 PM
>
> @TheLoneOlive American Airlines cnt fly individual drs/nurses in2 #haiti …Red Cross & other agencies are leading relief bit.ly/GIVE2HAITI

This was retweeted 14 times. In this case, using the hashtag #haiti was appropriate because the rumour was propagated through that hashtag. When countering a rumour, a person should use the same hashtag in which the rumour was propagated so that the same audience that may have been exposed to the rumour is now exposed to the counter argument. However, the addition of the hashtag #haitiearthquake would have increased the size of their audience and produced a more targeted crowd (people interested in the earthquake and not just Haiti).

Further, American Airlines use of "mention" is important to note. Although mentioning a person's twitter name is proactive and in line with a beacon model of rumour countering, @replying or mentioning a person's name does not mean that it will show up on the person's Twitter page. Only the person who was mentioned or @replied to will be able to see the @reply or mention. In other words, the person will be notified that American Airlines @replied to their tweet or mentioned them in a tweet. However, the only way other people or that person's followers will see the tweet is if they are following a hashtag included in the reply (i.e., #haiti), if they search on Twitter for words that are included in that tweet or if they search for that person's name. Although replying to a person's tweet by mentioning them or @replying to them is a more proactive and informal style which has many pros, it is important that the person or organization that is sending the counter message also uses the appropriate hashtag so that it is visible to a larger audience.

American Airlines was not the only organization to combat the Haiti rumour on Twitter. JetBlue also tried to counter the same rumour (Leberecht, 2010). Thus, we examined JetBlue's tweets from January 1, 2010 to March 1, 2010. JetBlue's first attempted to counter the rumour in the following way, twitter.com, 2015):

> JetBlue Airways @JetBlue 13 Jan 2010 7:40 PM
>
> @brooksbayne Please note: we're working with the Consulate and providing flights for some at their request, but not providing travel for all

This was retweeted only 4 times (@Brooksbayne account is protected and there is no access to his or her tweet). Then JetBlue followed up with the following tweet less than an hour later:

> JetBlue Airways @JetBLue 13 Jan 2010 8:25 PM

> Sorry if the last tweet sounded rude. We don’t have seats to accommodate the countless requests to assist in recovery efforts in Haiti

This was retweeted only 11 times and then they followed up with the following tweet a few minutes later:

> JetBlue Airways @JetBLue 13 Jan 2010 8:30 PM
>
> We wanted to address mis-information before it spread further. We’re utilizing available seats into SDQ for Haitian aid with the Consulate

This was retweeted only 14 times. JetBlue used no hashtags in their countering tweets. Also, in their first tweet they did not even use the word Haiti, nor did they use the word earthquake in any of their tweets. This might be the reason why their retweet count was so low.

Based on the results contained within this case study, we recommend that rumour countering strategies should aim to provide clear evidence for all rumour counters. We recommend the use of appropriate key words and appropriate hashtags to create a targeted audience.

1.2.3 Mumbai

Starting on November 26, 2008, ten terrorists attacked a number of buildings across Mumbai with automatic weapons and grenades, including the Taj Mahal Hotel, over the course of four days (“Mumbai terror attacks,” 2015). There were 164 fatalities and 308 people reporting injuries (Sabha, 2008). Ajmal Kasab (aka Mohammed Ajmal Amir) was the only surviving perpetrator. He was captured and jailed. Then on July 13, 2011, another terrorist attack occurred in Mumbai (Gupta & Kumaraguru, 2011), killing at least 21 people (Chakravorty & Jadhav, 2011). This time terrorists bombed the Opera House, the Zaveri Bazaar and Dadar (Gupta & Kumaraguru, 2011).

Rumours were spread during both the 2008 (Oh et al., 2013) and 2011 (Gupta & Kumaraguru, 2011) Mumbai attacks. Oh et al. (2013) found that approximately 35% of tweets related to the Mumbai terrorist attacks in 2008 were rumours. Gupta and Kumaraguru (2011) report that during the 2011 Mumbai terrorist attacks there “were false news, fake cries of help and negative sentiments that were propagated on Twitter” (p. 10).

The 2008 Mumbai attacks were considered different from typical terrorist attacks due to the high integration of Blackberries and GPS locators (Jenkins, 2009). This is because terrorists on the ground were communicating with another group of terrorists who were gathering information about the situation via social media. According to Oh et al (2011, p. 33), “situational information which was broadcast through live media and Twitter contributed to the terrorists’ decision making process and, as a result, increased the effectiveness of hand-held weapons to accomplish their terrorist goal.” Thus, this case study is an example of a tenuous situation where police have to weigh the importance of countering rumours with accurate information versus letting rumours spread. On the one hand, letting false rumours spread about the locations of police and other authorities may impede the terrorists’ actions. On the other hand, rumours can have unanticipated and significant impacts on innocent people and can impede the actions of police and medical personnel.

During the Mumbai attacks, the conditions were set for rumour generation. Oh et al. (2013) content coded tweets related to the 2008 Mumbai terrorist attack to examine the factors that best predicted whether or not a tweet would contain a rumour. The second greatest predictor of a rumour was personal involvement (Oh et al., 2013) or importance to oneself. Uncertainty was also a condition of the Mumbai attacks, which contributed to rumour generation. An analysis of both the 2008 (Oh et al., 2013) and 2011 (Gupta & Kumaraguru, 2011) Mumbai terrorist attacks suggest that there were a few differences in Twitter usage in the generation and propagation of rumours in comparison to analyses of rumour generation and spread in developed countries. First, unlike the UK riots, during the Mumbai terrorist attacks there was no official Mumbai Police Twitter account (Gupta & Kumaraguru, 2011). At the time, "developing nations like India, have only recently started becoming active on social media like Twitter. So there are only handful[s] of such authority users in India, who proactively tweet about or during incidents" (Gupta and Kumaraguru, 2011, p. 10). Further, Gupta and Kumaraguru (2011) found that highly followed accounts of government, media, and celebrities did not tweet very much. When they did, they were highly retweeted. Thus, we argue that the Mumbai attacks of 2008 and 2011 are a good example of an information vacuum. As such, it is likely that people did not have credible information from an authority during the Mumbai attacks. Indeed, "the majority of content generated at the time of crisis was from unknown users" (Gupta & Kumaraguru, 2011, p. 15).

Although Gupta and Kumaraguru (2011) did not compare rumour generation or propagation to developed countries, the absence of an authority presence on Twitter with the purpose of accurate information dissemination would likely lead to increased rumour spread via Twitter. A second difference noted by Gupta and Kumaraguru (2011) was that although the proportion of Twitter users tweeting URLs (40%) were similar to levels previously reported (high during crises; Hughes & Palen, 2009), the number of mentions were greater than levels reported by Hughes and Palen (2009). Whereas Hughes and Palen (2009) found that mentions decreased from approximately 22@ to 6-8% in emergency situations, Gupta and Kumaraguru found that approximately 28% of Twitter users who were tweeting about the Mumbai attacks used mentions. "A high number of @-mentions can be explained by the fact, that at times, like the Mumbai blasts, Twitter was being used as a medium by its users to exchange and coordinate information/news amongst each other, rather than its usual role as a micro-blogging forum to express one's personal opinion" (Gupta & Kumaraguru, 2011, p. 15).

Oh et al. (2013) found that source ambiguity (a lack of information about credibility or low credibility) was the greatest predictor of rumour. They coded a tweet as high on source ambiguity if it did not contain an external source such as name of media or links to media, video or picture or a tweet that "expresses distrust and or/ambiguity about the source" (Oh et al., 2013, p. 425). They provided the following tweet as an example of tweet low on source credibility because the person did not provide evidence for their claim.

> more hostages at the Cama hospital - #Mumbai

Source ambiguity was followed by indicators anxiety as a predictor of a rumour (Oh et al., 2013). This is in line with Oh and colleagues' (2010) previous finding that the level "a Twitter study of Mumbai terrorist attacks, online users' level of anxiety was not reduced over time, and hence, rumor mill never stopped throughout the twelve day terrorist attack" (Oh, Kwon, & Rao, 2010, p. 12).

In the absence of credible information, during the 2011 Mumbai terrorist attacks, Gupta and Kumaraguru (2011) identified the rumour that the terrorist attacks were perpetuated by the sympathizers of Ajmal Kasah (2008 Mumbai bombing captured terrorist) because the blasts occurred on his birthday. The rumour was only partly true, the blasts did occur on his birthday, but it was only a coincidence.

> Three blasts in Mumbai. It happens to be a birthday of the captured terrorist Kasah or was it planned to be on his birthday? #Mumbaiblasts.

This tweet was highly believable due to one aspect of the rumour being true and the highly coincidental nature of the rumour. Gupta and Kumaraguru (2011) found that this birthday rumour was tweeted 446 times and retweeted 223 times. Gupta and Kumaraguru (2011) found the first tweet of the birthday rumour (in their sample) occurred within an hour after the blasts. The tweets and retweets of this rumour decreased over time, and were very few at 24 hours after the rumour started.

Gupta and Kumaraguru (2011) also reported on another false rumour about a fourth blast (that did not occur). People even specified the location of the fourth attack. For example, a tweet taken from Gupta and Kumaraguru (2011) during the 2011 Mumbai attacks states:

> Fourth blast in Mumbai!!! At lemington road! Let da focus be on innocent ppl

This rumour was highly believable because of the specific location provided, even though any one can easily generate detailed, but false information.

Last, another tweet reported by Gupta and Kumaraguru (2011) during the 2011 Mumbai attack was a rumour that blood was required by hospitals. This resulted in many people turning up at hospitals and being turned away because their supply was full.

> RT idurgesh: #needhelp #MumbaiBlasts RT KapoorChetan: Bombay Hospital Blood Bank is in need of B+, B-, O+ blood groups. Please donate

This is a highly believable rumour because one might expect increased need for blood donations with the influx of patients to the hospital during a crisis situation, such as the Mumbai terror attacks.

1.2.4 The Boston Bombings

During the Boston Marathon on April 15, 2013, brothers Dzhokhar and Tamerlan Tsarnaev placed and detonated two pressure cooker bombs near the finish line. As a result, three people were killed and an estimated 264 others were injured (Kotz, 2013). Much discussion occurred via social media. Two sites in particular, Twitter and Reddit were an important medium in the propagation of at least four major rumours related to the Boston bombings (Starbird, Maddock, Orand, Achterman, & Mason, 2014; Gupta, Lamba & Kumaraguru, 2014; Maddock, Starbird, Al-Hassani, Sandoval, Orand, & Mason, 2015). The first rumour was the misidentification of Sunil Tripathi as a bombing suspect (Starbird et al., 2014; Maddock et al., 2015). The second rumour was that an 8 year old girl was killed while running the marathon (Starbird et al., 2014; Gupta et al., 2014; Maddock et al., 2015). The third rumour was that the U.S. Navy Seals were responsible for the bombing (Starbird et al., 2014; Maddock et al., 2015). And the fourth rumour was that a woman running the marathon was killed before a planned marriage proposal (Maddock et al., 2015). These rumours are of particular interest because both the public and officials were actively involved in the debunking process (primarily the Sunil Tripathi rumour).

Believability, credibility, and the social network all play a role in the propagation of rumours. And this can be magnified through social media. Rumours may have propagated to the extent that they did because they were believable and helped to make sense of the situation. For example, the rumour that a woman running the marathon was killed before her boyfriend could propose at the finish line was believable because finish line proposals are relatively common (Journal Sentinel, 2014) and it also serves to make sense of the question: “Who were the victims?” This was also true for the rumour that an eight year old who was running the marathon was killed. It helped put a face to the victims. The misidentification of Sunil Tripathi as a suspect served to answer the question: “Who did it?” That he could have detonated the bombs was believable because he had been reported missing a month prior to the bombings and was also reportedly struggling with depression (the link between mental health and violence is weak, but is nonetheless a pervasive stereotype; Friedman, 2006). Also, his skin and hair colour was a relatively close match to the suspects whose picture was released by the FBI. Finally, the Navy Seals rumour was believable, arguably to fewer people, due to pictures released that suggested that Navy Seals were there (Infowars.com, 2013).

Two of the rumours likely increased in believability once images were added to the text. For example, the first tweet of the girl ostensibly killed while running was only retweeted twice (Maddock et al., 2015). However, it emerged a second time, but this time it was attached to an image of a young girl running a 5K race (Maddock et al., 2015). In accordance with the definition of transmediation (i.e., take an image, change it and redistribute it; Bernardi et al., 2012), the rumour was appropriated (as opposed to retweeted which identifies the original source), changed (a picture was added), and retransmitted (retweeted). This addition of a picture likely lent credibility to the rumour, even though it was a picture of a 5K race and could have been easily taken from the Internet (which it was). Further, less than an hour after the rumour about the women who died before her boyfriend could propose spread through Twitter, users began to attach images to the tweet of a man treating an injured woman at the Boston marathon finish line (Maddock et al., 2015). Although the rumour was false, the picture was a real picture of the Boston Marathon bombings, appropriated from the Boston Globe without attribution and inappropriately attached to a false rumour. Maddock et al. (2015) found that the first burst of the rumour was primarily a text-based rumour, but the second burst consisted largely of variation of the original rumour linked to a photo. Maddock and colleagues suggest that the addition of the image catalyzed the second burst of the rumour through social media. Believability of a rumour increases when accompanied by attached photos. This seems to include "hard" evidence of the rumour's validity. And this can be accomplished more effectively using social media.

The Boston marathon rumours increased in propagation once the rumour was spread from credible sources (Gupta et al., 2014). Gupta et al. (2014) analyzed the propagation of a variety of Boston Marathon rumours on Twitter and compared them to true news. They found that the rumours had initial slow growth but once they were tweeted from a credible source, they went "viral" or experienced a very steep growth (calculated by tweets per minute). Gupta et al. suggest that it "may be attributed to the fact that the user profiles (source of a fake tweet) are people with low social status and unconfirmed identity. Hence, the initially fake tweet spread is slow, and they become highly viral only after some users with high reach (for e.g., large number of followers) propagate them further." (2014, p. 6).

Although the initial source of the Tripathi rumour is not easily determined, there were at least two (supposedly) credible sources that introduced Sunil Tripathi as a potential suspect: a previous classmate (Kundani, 2013) and a relative (Baijal, 2013). They both suggested that Tripathi matched the description (Kundani, 2013; Baijal, 2013). Once Tripathi's name was introduced to social media, Greg Hughes (an unofficial public source) tweeted that Tripathi was identified on police radio (Madrigal, 2013). Although this never occurred (Madrigal, 2013), this tweet was echoed by NBC and was associated with a rise in tweets (Maddock et al., 2015). Credibility of the rumour source then will likely increase the activity on social media regarding propagation.

On Reddit, there was a single crowd-sourcing community of armchair detectives within a subreddit/forum (www.reddit.com/r/findbostonbomers) which has since been taken down and that this is different than other types of social media because it is less segmented. According to DiFonzo (2008), the less segmented a population, the less diversity there is in a rumour. In line with groupthink (Janis, 1982) and group polarization (Isenberg, 1986) decreased diversity occurs in a population because people are exposed to global majorities and social pressures to submit (DiFonzo, 2008). The purpose of the subreddit r/findbostonbombers was "to mine through a massive amount of photos that had surfaced, including those posted on Flickr by Reddit users and published in a massive Google Doc titled 'Boston Bomber Info Spreadsheet'" (Abad-Santos, 2013, para 2). However, according to Abad-Santos, "the Reddit page that was so closely watched by reporters and social media users that it sparked digital witch hunts of innocent people" (2013, para 2). Unlike Facebook and Twitter and other social media, which are more segmented, everyone has a different social network, the subreddit r/findbostonbombers was the main, at least most popular, subreddit dedicated to the pursuit of identifying the suspects and this may have contributed to the pursuit of a single misidentified suspect, Sunil Tripathi. A lack of diversity on social media then can increase the propagation of rumours.

According to Starbird and colleagues (2014), the public was extremely involved in the countering of the Tripathi rumour and the "corrections persist[ed] long after the misinformation fade[d] as users commented on lessons learned about speculation." Starbird et al. (2014) suggest that this is likely due to the public's realization of their involvement in the generation and propagation of the rumour and the consequence of the rumour on Tripathi's family. Officials, such as the FBI, were indirectly involved in the debunking of the Tripathi rumour by releasing the names and pictures of the real suspects. But the purpose of which was likely more about public safety and the capture of the suspects than for the purpose of debunking the rumour of Tripathi as a suspect. Further, we found that the FBI did not communicate information via Twitter or other social media during the bombings, but through official press releases that were picked up by the media. Once the FBI notified the public of the true suspects, the public was actively involved in the dissemination of this information through social media.

The FBI was likely perceived as a credible source and most knowledgeable on matters of national security. When the true names of the suspects were released there was little doubt in the public that these were the actual suspects. The suspect information was disseminated by mainstream media sources, also likely perceived as credible. The dissemination of this information from credible media sources (i.e., NBC and Associated Press) was associated with a decreased in the Tripathi rumour propagation on Twitter (Maddock et al., 2015).

When the FBI and local police forces released the names of the suspects, they also released their pictures, which is clear evidence that can be shared via social network. These pictures showed the suspects at the Boston marathon with backpacks. Further, when Dzhokhar Tsarnaev was arrested, a member of the Massachusetts State Police provided to Boston Magazine photos of Tsarnaev captured which were subsequently released (Memmott, 2013). These pictures provided clear evidence that Tsarnaev was the suspect, and not Tripathi.

The rumour of the death of the eight year old girl was countered by the mainstream media in the releasing of names and descriptions of the real victims and this information was disseminated by the public (Starbird et al., 2014). The FBI, mainstream media and the public countered the Tripathi and the eight year old girl victim rumour with direct information, i.e., the release of the names and descriptions of the real suspects and the real victims. Facts then debunked the rumour.

Although most of the Boston bombing rumours were countered with direct information, a user-based denial-of-service method was also used to counter the Tripathi rumour. On reddit.com, where the rumour of Sunil Tripathi was highly propagated (Starbird et al., 2014; Maddock et al., 2015), once the FBI named the official suspects, moderators of the r/findbostonbombers subreddit prevented users from identifying any other suspects other than the two official suspects by deleting any posts pointing to different suspects (Kleinman, 2013).

Timing is also a critical strategy for debunking rumours. The eight year old female victim and proposal rumours were debunked once the media released the names and descriptions of the actual victims (Starbird et al., 2014). It became clear that it was an eight year old male spectator, not a female runner who was killed and it was a young woman spectator who was killed and not a young woman marathon runner. However, the Tripathi rumour continued to propagate online for three days (Starbird et al., 2014). In alignment with the delayed start model, the FBI did not provide any information on the bombing suspects until three days after the bombing. Reasons for this are unknown. However, it is likely because it took time to determine, verify and issue a warrant for the suspects. In the information vacuum, rumours were generated to make sense of the situation.

Although the Navy Seals rumour was not countered by the FBI or other official source, it would have been a difficult rumour to counter. The Navy Seals rumour is a conspiracy rumour which is harder to prove because at the heart of the rumour is a coverup, so any debunking information released by an official source would likely have been perceived as a coverup by those who believed in this rumour and may have even increased their belief in the rumour. Leaving it alone might be the most effective strategy to debunk this conspiracy rumour.

1.2.5 Noordin Mohammed Top

Noordin Mohammed Top was an Islamist terrorist, Indonesia's most wanted Islamist militant, who had been assumed killed in a raid by the Indonesian police on September 17, 2009 (Bernardi et al., 2012). Instead, according to Bernardi et al. (2012), he lived and achieved a sort of mythical status among the Jemaah Islamiyah (JI) terrorist group. One and a half months later, he was officially killed by the Indonesian police. Shortly after his death, an autopsy was conducted. Based on the results of the autopsy, a forensics expert and police spokesman announced to the media that it was likely that Noordin Mohammed Top was gay. After the announcement on mainstream media, the "rumour" spread like wildfire online and through social media. The rumour was particularly image driven, with his image manipulated to look effeminate on pictures, videos and mashups ("recombination of graphics, image, and text from various sources to create a derivative work" Bernardi et al., 2012, p. 167).

This rumour would be classified as a hostile rumour due to the decreased acceptance of homosexuality in Indonesia, and particularly in groups like the JI that shuns homosexuality (Bernardi et al., 2012). According to Bernardi et al. (2012, p. 133), "the Noordin story appears to have been a successful instance of rumour employed as an information warfare countermeasures, primarily because of the nature of the origin of the Noordin rumour as well as its titillating nature, which allowed it to become a meme, gain traction, and spread across Indonesia and the region."

The rumour that Noordin Mohammed Top was gay propagated wildly through Indonesian social media. "Noordin M. Top" was the most searched term on Yahoo in Indonesia in 2009 (Bernardi et al., 2012). Two spikes in viewing occurred. First, after Top was officially killed and another during the press conference releasing details from the autopsy. Bernardi et al. (2012, p. 130) suggest that "the frequency of repetition on various media such as YouTube, Facebook, and blogs heightened public awareness of this rumour beyond the more mainstream news media representations." This frequency with which the rumour appeared may have increased its believability, as frequency and believability are likely positively correlated. The rumour was also emotionally charged and thus more likely to be propagated.

The propagation of the Top rumour through social media is an excellent example of the transmediation and presumption processes. The Top rumour originally spread as reposting of the original media stories, then the commentaries attached to these news stories turned into parody and crude jokes. Then bloggers took official images and manipulated them by giving him a jilbab, added long hair, coloured his cheeks or made him look like Michael Jackson (Bernardi et al., 2012). The ability to modify images and share them across social media helps to propagate rumours.

Although the effectiveness of this rumour cannot be directly examined, there are some indicators that suggest that this rumour helped to discredit a glorified terrorist. First, Bernardi et al. (2012, p. 129) suggest that the rumour had the potential to be powerful: "humiliation is a powerful weapon in the war against Islamist terrorism. … Osama bin Laden feared humiliation more than death." Second, the rumour had the potential to discredit (in the eyes of other Islamist extremists) the Islamist extremists that he was associated with. According to Bernardi et al. (2012, p. 123), the rumour of his homosexuality "cast doubt among Islamist extremists and potential converts about his piety and suitability as an Islamic martyr…Noordin was a very important figurehead for the organization; this rumour effectively painted him as a hypocrite and thus, by proxy, the organization as hypocritical." Third, there is some evidence that the rumour impacted Islamist extremists such that they disassociated or at least, did not glorify a deceased terrorist. As Bernardi et al. (2012, p. 131) state,

> "It is striking to note that, unlike other terrorist leaders who have garnered posthumous acclaim, Noordin seems to have receded in discourse after the release of the rumour. His image was not enhanced with symbols normally associated with martyrdom in the Muslim tradition. [Unlike how fans mourn their] heroes and idols collectively on mediated platforms, which then further inflames their passion and resurrects the fame of deceased celebrities. In Noordin's case, online discourse appeared not to have any such vivifying effects."

Fourth, the rumour was difficult to refute, "if not impossible, for defenders of Noordin to dispel without access to the corpse or scientific debate" (Bernardi et al., 2012, p. 131). Indeed, this rumour seemed to overpower the statement from Noordin's family who denied the allegations. Bernardi et al. (2012) suggest that the rumour campaign spread through social media successfully undermined Top's legacy as a martyr.

As stated by Bernardi et al. (2012, p. 127), "the rumor's primary purpose was to defuse extremist rhetoric and the valorization of Noordin among the population as a devout Muslim and a martyr." For a rumour to take off, it needs to be a rumour that resonates with the people and their culture (Bernardi et al., 2012), and this rumour did that successfully.

1.3 Recommendations

Based on the review, we recommend the following ways to counter rumours: 1) early identification and refutation; 2) high source credibility; 3) clear and strong evidence; and 4) ensuring a captive audience. We also have some specific recommendations for the use of social media in rumour identification and countering. First, during a crisis situation it is important to keep track of "local knowledge gatherers" whether this is a group created to disseminate situation information or a crowdsourcing group of armchair detectives. We recommend two specific strategies on twitter, creating the use of #rumour hashtag and monitoring one's mentions and @replies in order to keep tabs on the development of rumours and to be able to quickly respond to rumour-related inquires.

In terms of rumour countering, we recommend disseminating the refutation to an audience that has a large and targeted audience. Specifically, we recommend mentioning relevant influential (i.e., high number of followers) Twitter users in the counter to increase the chance that they will retweet one's message. We also recommend directly responding to users, particularly those who propagated a rumour. Lastly, we recommend that when using hashtags that one considers their audience and the type of audience who will be following or not following certain hashtags. Further, using the appropriate key words is also essential in order for the message to be received by a relevant audience.

We also recommend that when directly refuting a rumour is not possible or its predicted efficaciousness low, consider using a counter rumour to fight rumours with rumours by either providing counter information in the form of a sensational rumour or by discrediting the source of a rumour.

Finally, we recommend that rumour countering should be considered from a strategic and tactical lens. Countering a rumour may require the release of classified information. Not countering a rumour may lead to a negative impact on the public or a strategic military mission. In other words, the consequences of countering a rumour or allowing a rumour to propagate should be carefully considered.

Overall, this report demonstrates that understanding the details of each social media platform and understanding the implications these platforms have for rumour generation, propagation and countering is essential to effectively prevent rumour spread in crisis, war or politically sensitive situation.

1.3.1 Future Research

It is important to note that the effectiveness of countering strategies used in the following case studies was difficult to measure and either was not attempted by researchers or was hedged. The largest issue related to measurement of effectiveness is that rumours subside over time (Shibutani, 1966), thus for any rumour countering strategy to be deemed effective it must demonstrate a greater or quicker decrease in rumour propagation than would naturally occur. Because a control group or a control rumour is difficult to find in an applied setting, rumour countering strategies were primarily described as opposed to assessed in terms of effectiveness. A properly designed applied research study adopting a quasi-experimental method aimed at examining the impact of a rumour countering strategy in the real world could more rigorously determine its effectiveness. The details of this design are discussed in more detail in the section on future research.

It is apparent that there is need for a rigorously designed, yet applied, quasi-experimental study of rumour countering on social media in real time. Because we cannot randomly assign people to hear a countering message or control message in real life, we have to choose groups of people who have vastly different online social networks but who are similar nonetheless and expose them to a counter versus control message and then assess the impact of the counter message. Thus, one can assess the impact of a counter message on the same social media interface and compare it to how the rumour would naturally diminish over time. A number of controls would have to be implemented to prevent the control group from receiving the counter message.

One very important factor in rumour countering that has become apparent through the literature review and case studies is early refutation. In order to counter a rumour in its early stages, one must have also identified it early. To this end, a number of small scale investigations have examined or proposed ways to identify rumours as they occur in real time (Mendoza, Poblete, & Castillo, 2010; Qazvinian, Rosengren, Radev, & Mei, 2011; Seo, Mohapatra, & Abdelzaher, 2012; Ratkiewicz, Conover, Meiss, Goncalves, Patil, Fammini, & Menczer, 2010; Ennals, Byler, Agosta, & Rosario, 2010). A review of these findings may be important as well as a large scale study on rumour identification.

Social media may not have changed the way in which rumours recur, but it has changed the ease in which we can track and analyze rumours. For example, Twitter data can be analyzed with Streaming APIs which provide open access to tweet data for developers and researchers. It can help researchers collect a sample of tweets from which they can estimate the number of tweets per topic and assess the number of tweets per topic per timeframe (e.g., Tweets of certain rumour per minute) as well a variety of other indicators (geo data, etc.). Integrating this into an applied study might be an effective way to measure rumours and social media more systematically.

This page intentionally left blank.

2. How a High Status Outgroup can Influence a Low Status Ingroup

In the following section, we seek to understand how groups can influence each other. More specifically, we review the social identity theory literature to identify effective techniques that outgroup members can use to influence the core beliefs and attitudes of a target group.

2.1 Resistance to Criticism

Group criticism refers to the analysis and judgment of a particular group based on perceived faults (Oxford, 2015). Criticism is an important tool for encouraging groups to change their attitudes and behaviour as it highlights areas for improvement (Rabinovich & Morton, 2010). At the same time, criticism can be a “missed opportunity” for change as groups tend to ignore negative information about themselves when this information is given to them by outsiders (Hornsey & Esposo, 2009). This phenomenon is referred to as the intergroup sensitivity effect (ISE) and describes the tendency of individuals to react more defensively to criticism made by an outgroup member compared to an ingroup member even when the message is identical (Hornsey & Esposo, 2009). For example, individuals agree with criticism to a lesser extent and view the critic’s comments more negatively when the critic belongs to an outgroup than when he or she belongs to an ingroup (Brander & Hornsey, 2006; Hornsey, Robson, Smith, Esposo, & Sutton, 2008).

Defensive reactions to criticism, such as rejecting the criticism, occur because individuals are inclined to believe that outgroup critics have ulterior motives (Hornsey & Esposo, 2009). Whereas ingroup critics are assumed to have helpful motives, such as concern about the group’s welfare, outgroup critics are thought to have destructive motives, such as a desire to overpower or demoralize the group (Judd, Park, Yzerbyt, Gordijn, & Muller, 2005, as cited in Hornsey et al., 2008; Rabinovich & Morton, 2010; Tanis & Postmes, 2005, as cited in Hornsey et al., 2008). As a result, individuals are more likely to ignore negative feedback given by outsiders, but accept it when it is given by ingroup members. In support of this, research has found that individuals attribute more constructive intentions to ingroup critics than they do to outgroup critics, and this bias statistically accounts for the intergroup sensitivity effect (Hornsey & Esposo, 2009).

Researchers have found that the same defensive reactions occur if the ingroup is skeptical of another ingroup member’s motives when he or she is delivering criticism. More specifically, if individuals are skeptical of the ingroup critic’s motives because he or she is not committed to the group (Hornsey, Trembath, & Gunthorpe, 2004) or he or she is a newcomer (Hornsey, Grice, Jetten, Paulsen, & Callan, 2007, as cited in Hornsey et al., 2008), the ingroup reacts as defensively as they would to an outgroup critic. In other words, individuals react defensively to criticism about their ingroup when the critic is seen as having destructive motives – regardless of whether the critic belongs to the ingroup or an outgroup. Overall, this research suggests that attributions about constructive versus destructive motives explain why the intergroup sensitivity effect occurs.

Defensive reactions to criticisms can have important implications for intergroup relations. For example, outgroup critics arouse negative affect and are considered undesirable interaction partners (Hornsey & Esposo, 2009). These negative perceptions can spill over to the critic’s entire outgroup, creating tenuous relations and inciting conflict (Branscombe, Ellemers, Spears, & Doosje, 1999, as cited in Hornsey & Esposo, 2009). Thus, rather than promoting change, criticism may make change *less* likely while also eroding intergroup trust (Hornsey & Esposo, 2009).

Given the potential of outgroup criticism to hinder intergroup relations, it is important to understand how to override the intergroup sensitivity effect. This can be done by using strategies that target the underlying mechanism of the effect. Thus, any strategy that alters the ingroup's perceptions of an outgroup critic's motives should be effective in increasing feedback acceptance and yielding attitude change. In the sections below, we review strategies that have been shown to be effective in reducing defensive reactions to outgroup criticism while also highlighting strategies proven to be ineffective. We also explore additional mechanisms that explain why the intergroup sensitivity effect occurs.

2.2 Effective strategies for removing resistance to criticism

2.2.1 Praise should accompany criticism

Hornsey and colleagues (e.g., Hornsey & Esposo, 2009; Hornsey et al., 2008) suggest that attaching praise to criticism may reduce defensive reactions, such as disagreeing with the criticism and viewing it negatively. They argue that using praise suggests that the outgroup critic is motivated by constructive reasons and consequently, increases the probability that ingroup members will process the critic's message and incorporate his or her perspective (Esposo, Hornsey, & Spoor, 2013; Hornsey et al., 2008). An alternative perspective is that ingroups may view praise as an attempt to manipulate or deceive the ingroup; if this were the case, ingroups would react defensively by rejecting the critic's message. In fact, to the extent that negative feedback overshadows positive feedback, attaching praise to criticism may have little impact (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001, as cited in Hornsey et al., 2008).

Hornsey et al. (2008) explored these possibilities in an experiment in which 107 Australian undergraduate participants read an excerpt about the racism of their ingroup that was supposedly written by another Australian (ingroup) or a non-Australian (outgroup). The criticism was either accompanied by praise, highlighting what the critic liked about Australians, or it was not (control group). The results indicated that individuals who encountered praise from either an ingroup or outgroup critic attributed more constructive motives to the critic, liked the critic more, and indicated stronger agreement with the critic's message. However, praise had no effect on participants' feelings of negativity (e.g., irritated, offended) towards the critic's comments: participants felt equally negative about the criticism regardless of whether or not praise was used. Importantly, these effects occurred for both ingroup and outgroup critics, suggesting that this strategy is not specific to the intergroup sensitivity effect and is more broadly applicable.

These results suggest that attaching praise to criticism may be an effective strategy for reducing defensive reactions and increasing attitude change. The results also support Hornsey et al.'s (2008) argument that "sugar coating" criticism is effective because "nice" critics are attributed more constructive motives. Indeed, they found that perceptions of constructiveness mediated the relation between praise and attitudes (i.e., agreement, liking). In other words, participants attributed more constructive motives to critics who used praise compared to those who did not and these perceptions led to more positive reactions, such as liking the critic more. Moreover, this effect occurred even though participants generally attended to criticism more so than praise, suggesting that exposure to negative information does not overpower the impact of positive information in this context.

Despite the benefits of this research, it is important to note that although using praise helped to increase the likeability of the critic and agreement with his or her message, it still resulted in negative emotional reactions. This indicates that praise is not effective in reducing all types of defensive reactions (Esposo et al., 2013). Finally, because the results of one study are not definitive, more studies are needed to demonstrate the effect's replicability.

2.2.2 Own-group criticism

Another strategy that outgroup critics can use to defuse suspicion about motives is to acknowledge their own group's failings (Hornsey et al., 2008). Ingroups commonly assume that outgroup members use criticism to demoralize the group and/or to achieve supremacy over that group. However, acknowledging one's own group's failings works against the goal of achieving group domination because it exposes the group's weaknesses, making it vulnerable. Own-group criticism may act to diffuse suspicions that the outgroup critic has malevolent intentions and increase the probability of message acceptance.

In one experiment, Hornsey et al. (2008) investigated the extent to which an outgroup member who simultaneously criticized his own and another group created the perspective that he or she is concerned about the ingroup's welfare. One-hundred-and-sixty Australian undergraduate students read an excerpt written by a member of their ingroup or by a member of their outgroup in which the critic disparaged Australians for being racist. In addition, the outgroup critic acknowledged racism within his or her own country, made no acknowledgment, or acknowledged racism within other countries (to ensure that any effects were due to own-group criticism and not to acknowledgement in general). Next, participants completed measures of critic likeability, agreement with the message, the constructiveness of the critic's comments, and feelings of negativity towards the critic's comments (e.g., offended, irritated). The researchers also measured perceptions of hypocrisy to ensure that any positive effects of own-group criticism were not due to the fact that the critic was simply seen as being less hypocritical. The results of the experiment are shown in the table below for ease of interpretation.

Table 1: Effects of criticism on reactions (in Hornsey, Robson, Smith, Esposo, & Sutton, 2008)

		Outgroup Critic		
		Acknowledgement of Group Failings		
	Ingroup Critic	Own Group	Other Group	None
Likeability	4.18_a	4.13_a	3.60_b	3.46_b
Agreement	4.25_a	3.97_a	3.66_a	3.78_a
Constructiveness	4.32_a	3.98_a	3.31_b	2.69_c
Negativity	3.71_a	4.18_{ab}	4.69_b	5.31_c

Note: Cell entries represent means.

Means that do not share the same subscript are significantly different from one another. Specifically, the results indicated that critic likeability and perceptions of constructiveness were highest and feelings of negativity were lowest when the criticism was made by an ingroup member *and* when an outgroup critic acknowledged failings of his or her own group. Moreover, this effect was mediated by perceptions of constructiveness such that participants' reactions to criticism were determined by how constructive they thought the critic's motives were – more constructive motives elicited more positive reactions. In addition, hypocrisy was not found to mediate this effect, suggesting that it is not a competing explanation.

Overall, the results of Hornsey et al.'s (2008) study indicate that one way outgroups can mitigate an ingroup's defensive reactions to criticism is to acknowledge the failings of their own group on the same dimension. However, it is important to note that although this strategy was successful in eliminating the intergroup sensitivity effect on measures such as critic likeability, its effect on agreement is unclear. That is, participants agreed with the message equally regardless of condition. This contradicts previous research in which individuals in the ingroup critic condition agreed with the critic's message significantly more than individuals in the outgroup condition. The extent to which this strategy is effective at persuading ingroup members to change their views is unclear. Attempting to replicate these findings in additional studies should clarify this issue. Finally, their study only investigated acknowledging one's group's failure on a factor directly related to the criticism (e.g., racism). It is unclear whether or not these findings would generalize to situations in which outgroup critics acknowledge the shortcomings of their group on factors unrelated to the matter at hand. Nevertheless, Hornsey et al.'s (2008) study demonstrates that own-group criticism is effective because it leads to the perception that outgroup critics have the ingroup's best interest at heart. It would appear to emphasize constructive criticism over destructive criticism. Similarly, Saguy and Halperin (2014) argue that own-group criticism is effective because it suggests that outgroup members are willing to consider the ingroup's perspective in order to resolve intergroup differences When in conflict, groups may assume that the other group is biased and rigid in their views (Saguy & Halperin, 2014). However, exposure to an outgroup member who acknowledges the flaws of his or her own group suggests that the outgroup is willing to consider the ingroup's perspective in order to reach a compromise. This gesture may make ingroup members more hopeful about the future of the conflict (i.e., because there is evidence that the outgroup is willing to compromise) and lead them to reciprocate by listening to the outgroup's perspective (Saguy & Halperin, 2014).

In a series of studies, Saguy and Halperin (2014) investigated the role of own-group criticism in increasing acceptance of the outgroup's perspective in the context of the Israeli–Palestinian conflict. This conflict has been described as hopeless because neither partner seems willing to work collaboratively (Halperin & Bar-Tal, 2007, as cited in Saguy & Halperin, 2014). Saguy and Halperin (2014) argue that believing that the other party is willing to consider a balanced perspective, as indicated by own-group criticism, may make the ingroup more willing to consider the outgroup's divergent perspective and to change their attitudes. In the first study, 91 Jewish Israeli university students read a fake UN report on the Israeli-Palestinian conflict. The own-group criticism condition included an excerpt by a Palestinian official derogating the Palestinians for their use of violence, whereas the control condition did not. Participants subsequently completed measures of hope regarding the Israeli–Palestinian conflict and openness to other group's perspective (e.g., reading books that represent the Palestinian point of view). The authors hypothesized that exposure to own-group criticism would make the ingroup more hopeful about the future of the conflict and, in turn, lead them to be more open to hearing the outgroup's perspective compared to the control condition.

Consistent with their hypotheses, the results from Saguy and Halperin (2014) indicated that individuals exposed to an outgroup critic expressing own-group criticism were more hopeful about the future of the conflict and more open to hearing the other group's perspective compared to the control group. Moreover, the effect of condition on openness was partially mediated by hope. In other words, individuals were more open to considering the outgroup's perspective because they were more hopeful about the outgroup's desire to collaborate in resolving the conflict. One limitation of their first study is that participants may have been more open to the Palestinians' perspective because they believed that the Palestinians shared their point of view. To counter this, the second study examined own-group criticism for an issue that was unrelated to the conflict – the Palestinian education system (for which Israelis should not have a pre-existing opinion). Using a different scenario than and the same measures as the first study, the researchers replicated the results of study 1, suggesting that exposure to own-group criticism may make ingroups more open to listening to the outgroup's perspective even if it is different from their own. In a subsequent experiment, Saguy and Halperin (2014) sought to further explore why own-group criticism engenders hope and leads to openness. The authors hypothesized that own-group criticism would lead to the belief that the outgroup is an open-minded partner, which would in turn foster more hope regarding the future of the conflict. Moreover, hope would make ingroup members more inclined to consider the Palesentians' point of view, which would then lead to more political compromise. Ninety-six Jewish Israelis recruited online participated in this study. Individuals in the criticism and control group read the same scenario as in study 1. Saguy and Halperin included measures of hope, the belief that the outgroup is an open-minded partner, openness to hearing the outgroup's perspective (e.g., "meeting personally with Palestinians to hear their point of view of the conflict, *even if that point of view is opposite to yours*"), and political compromise (e.g., "Given a peace agreement, to what extent do you support Israel taking partial responsibility for the refugee problem by allowing a limited number of refugees to enter Israel?")

The results in Saguy and Halperin's (2014) study demonstrated that individuals in the own-group criticism condition were more likely to view the Palestinians as open-minded, were more hopeful about the future of the conflict, and were more interested in hearing the Palestinians' perspective compared to the control condition. Although participants in both conditions were reluctant to engage in a political compromise, they were more inclined to do so in the own-group criticism condition.

Overall, Saguy and Halperin (2014) found support for their hypotheses that own-group criticism leads to increased hope through the mediating role of outgroup open-mindedness. Moreover, the effect of outgroup criticism on willingness to hear the outgroup's perspective was mediated by outgroup open-mindedness and hope. Openness to the Palestinians' perspective led to more political compromise.

This research suggests that by acknowledging the failings of one's group, outgroup critics can effectively persuade ingroups to change their attitudes and political views in addition to making them more inclined to listen to their perspective. Moreover, by exploring this research question using a highly salient and relevant scenario (i.e., Israeli-Palestinian conflict), we can be more confident that these findings will generalize to real-world scenarios.

2.2.3 Attributions for behaviour

Ingroups react defensively to criticism because they make attributions about the outgroup critic's motives (Esposo et al., 2013; Hornsey, 2005). However, the tendency to make attributions is a "two-way street." When criticizing other groups, critics almost always make attributions as to why a group may have acted in a certain way. Rabinovich and Morton (2010) argue that the effectiveness of criticism is contingent on the *kind* of attribution the critic makes about the group's behaviour. More specifically, they argue that individuals are more likely to change their behaviour in response to an outgroup critic when the critic attributes the group's flaws to internal features of the group, such as personality or ability, than when he or she attributes flaws to external factors, such as situational constraints (e.g., attributing poor performance on a test to lack of intelligence is making an internal attribution, whereas attributing poor performance to test difficulty is an external attribution). Internal attributions for group failure by outsiders threaten the group's positive image (Weiner, 2001, as cited in Rabinovich & Morton, 2010).

For example, suggesting that the Canadian government's decision not to send Stephen Harper to the UN climate summit reflects Canadians' apathy towards environmental issues threatens our image as a socially responsible nation. As a result of these attributions, ingroups may be more motivated to change their behaviour in order to restore their group's positive image (Rabinovich & Morton, 2010). Canadians might invest more time demonstrating that they are environmentally conscientious. On the other hand, external attributions for failure do little to threaten the group's image, because factors unrelated to the group are used to justify poor behaviour (Gold & Weiner, 2000, as cited in Rabinovich & Morton, 2010).

Research found that ingroup members' intentions to change their behaviour increased as a function of how threatening the internal attributions were (Iyer, Schmader, & Lickel, 2007). Individuals were motivated to change their behaviour to improve the group's image as a result of feeling shame or anger about the ingroup's behaviour (Iyer et al., 2007). In contrast, internal and external attributions by an ingroup member are equally non-threatening. Again, with respect to external attributions, factors outside of the group's identity are used to explain failure, suggesting that it is situational factors that need to change and not the group per se. With respect to internal attributions for failure, although they do acknowledge the group's deficiencies, as long as they are not shared with outsiders, the group's image remains intact, thus failing to incite need for change (Rabinovich & Morton, 2010).

Other research has found that ingroup members who criticize their own group are viewed much more negatively when they express these views to other groups than when they express them internally (Hornsey, De Bruijn, Creed, Allen, Ariyanto, & Svensson, 2005). Thus, as long as criticism is kept "in-house", it is not perceived as threatening. This research may seem at odds with Hornsey and colleagues' (e.g., Hornsey & Esposo, 2009) argument that ingroups react defensively to outgroup criticism as a result of group image concerns. That is, ingroups disagree with the criticism and view the critic negatively when they believe the critic seeks to hurt the ingroup's image.

Where most research on the intergroup sensitivity effect has focused on ingroup attitudes (e.g., feelings of negativity), some research focuses on *behavioural* responses to criticism. It is possible that although ingroup members may not agree with the criticism and feel negatively about it, they may still change their behaviour in order to repair their group's image.

In two experiments conducted by Rabinovich and Morton (2010), university students were given a fake newspaper article criticizing the UK's lack of adherence to recycling norms. The commentary was delivered by either a British (ingroup) or European (outgroup) government representative. The spokesperson either attributed the UK's behaviour to external reasons – the lack of infrastructure to support recycling – or internal reasons – UK citizens are irresponsible. After reading the article, participants indicated their intentions to engage in specific environmental behaviours, such as donating money to an environmental cause. The results consistently demonstrated that participants exposed to the outgroup critic intended to behave more environmentally-friendly when the critic attributed poor performance to internal rather than to external reasons. On the other hand, participants who read a message from an ingroup critic did not behave differently when the critic made internal attributions compared to when he or she made external attributions. In a third experiment, Rabinovich and Morton (2010) found support for the argument that this effect occurs because ingroup members want outgroups to view them in a positive light. More specifically, when internal attributions for behaviours were made, participants indicated higher intentions to behave environmentally-friendly when they believed that an outgroup member would scrutinize their responses than when an ingroup member would. On the other hand, when external attributions were made, believing that an outgroup member would see their responses elicited the same behavioural intentions as believing an ingroup member would.

Overall, this line of research suggests that criticism accompanied by internal attributions for an ingroup's behaviour may be an effective means of motivating behavioural change when the criticism is delivered by an outsider. Importantly, however, these results do not address the extent to which this technique affects ingroup attitudes. It is possible that even though ingroup members were motivated to change their behaviour, they may have still felt negatively about the criticism and the outgroup. Thus, it is not clear how this strategy would affect intergroup relations or conflict resolution.

2.2.4 Invoking a Common Identity

Individuals react defensively to criticism when the critic is assumed to have malevolent intentions regardless of whether the critic belongs to the ingroup or an outgroup (Hornsey, 2005). The extent to which a critic is perceived as having deconstructive motives is partially contingent on perceptions regarding his or her identification with the ingroup. Critics who identify weakly with the ingroup are assumed to have self-serving motives whereas critics with strong identification are assumed to have helpful motives (Hornsey et al., 2004). Past research has shown that ingroup members react defensively to low identifiers, but not to high identifiers (Hornsey et al., 2004). Thus, one effective strategy that outgroup critics can use to reduce an ingroup's defensive reactions to criticism is to make their identification with the ingroup salient.

Hornsey and colleagues (e.g., Hornsey, 2005; Hornsey et al., 2004) argue that using inclusive language is a powerful means of situating oneself within a group. For example, stating "Us Canadians are politically apathetic'' conveys a sense of commitment to and identification with the group, whereas "Those Canadians are politically apathetic'' conveys a sense of psychological detachment from the group (Hornsey, 2005). Although both statements reflect the same criticism, the first is likely to be more effective as this critic will be perceived as being committed to the group and, consequently, as having more constructive motives. Thus, when the outgroup critic makes claims to insider status by using inclusive language (Hornsey, 2005), he or she can accrue the same benefits that are afforded to ingroup members. This is consistent with Gaertner and Dovidio's (2000, as cited in Hornsey et al., 2004) common ingroup identity model in which intergroup conflict is reduced by creating the realization that both groups share an overarching identity.

To test the effects of inclusive language on the intergroup sensitivity effect, Hornsey et al. (2004) recruited 115 Australian undergraduate students to participate in an experiment. Participants read an excerpt by either an Anglo- or Asian-Australian that criticized Australians. The individual either used inclusive (e.g., "we") or exclusive language (e.g., "they"). The researchers hypothesized that participants would react more defensively to critics who used exclusive language compared to critics who used inclusive language. Consistent with these hypotheses, the results indicated that critics who used inclusive language were rated as being more attached to the Australian identity and were given more favourable trait evaluations (e.g., intelligent, trustworthy) compared to critics who used more exclusive language. Moreover, participants rated comments using inclusive language as more constructive and less negative (e.g., offensive) than comments using exclusive language. Contrary to expectations, however, using inclusive language did not make participants more likely to agree with the critic's message.

Overall, these results suggest that invoking a common identity by using inclusive language may be an effective means of reducing defensive reactions to criticism. It is important to note that, although inclusive language elicited positive effects on feelings of negativity, for example, it was not effective in eliciting attitude change. An important caveat of this strategy is that it cannot be used by anyone. For example, a male Caucasian cannot simply use the term "we" when talking about female Hispanics. Instead, the outgroup critic and ingroup must also belong to a similar superordinate category (e.g., Canadians; Hornsey et al., 2004). In fact, in Hornsey et al.'s (2004) study, although both critics had different ancestry (i.e., Anglo versus Asian), they were both Australian. Using inclusive language is just one method individuals can use to situate themselves within a group.

Schmader, Croft, Whitehead, and Stone (2013) suggest that emphasizing collective strengths or group-based needs is another means of invoking common identity. Highlighting collective strengths suggests that one strongly identifies with the group and places the group's needs before personal needs. In contrast, highlighting personal strengths suggests that one is more committed to oneself than to the group. Schmader et al. investigated this strategy within the context of prejudice in personnel selection.

They hypothesized that stigmatized individuals who referred to collective strengths in their application would experience no discrimination, whereas those who referred to personal strengths in their application would experience discrimination. In one experiment, American participants played the role of an interviewer in which they were asked to select two of three applicants for a job interview: a white heterosexual male candidate with weak qualifications, a white homosexual male candidate with average qualifications, and a white heterosexual male candidate with strong qualifications. They manipulated common identity by having the gay candidate emphasize collective strengths (i.e., a common identity) or personal strengths (i.e., self-promotion). For example, in the common identity condition, the candidate referred to the superordinate identity of being American, indicated his preference for team work, and using the pronoun "we." In the self-promotion condition, the applicant referred to his personal strengths and used the pronoun "I."

The results indicated that whereas all participants selected the highly qualified heterosexual white male, deciding whether to select the gay average candidate or the heterosexual weak candidate was contingent on whether or not the gay applicant invoked a common identity. Specifically, when the gay candidate emphasized personal strengths, participants were equally likely to choose him and the heterosexual weak applicant, even though the gay applicant was actually a better performer. On the other hand, when the gay candidate emphasized collective strengths, participants were significantly more likely to select him over the weak candidate. Schmader et al. (2013) also found that whereas participants in the self-promotion condition viewed the gay average applicant and the heterosexual weak candidate as being equally warm and competent, participants in the common identity condition viewed the gay applicant as being significantly more warm and competent than the weak candidate.

Though the preceding results do not speak directly to the intergroup sensitivity effect, they suggest that stigmatized individuals, such as outgroup critics, are more likely to elicit positive reactions from others when they invoke a common identity. Invoking a common identity can be achieved by directly referring to a superordinate identity with which the ingroup identifies (e.g., Canadians) or by using inclusive language.

2.3 Ineffective strategies for removing resistance to criticism

Strategies that fail to target the underlying mechanism(s) driving the intergroup sensitivity effect tend to be ineffective at reducing resistance to criticism. In the section below, we review some of these ineffective strategies and explain why these strategies are ineffective.

2.3.1 Argument Quality

Theory and research suggest that when individuals systematically process information, they are more persuaded by strong arguments than by weak arguments (Petty & Cacioppo, 1986). However, when individuals choose not to process information systematically, such as when they are uninterested or unmotivated, they are more likely to rely on superficial cues when evaluating argument quality, such as cues relating to the source of the message (e.g., credibility, attractiveness).

Building on these ideas, Esposo et al. (2013) proposed a model that describes the effect of argument quality on attitude change in the context of the intergroup sensitivity effect. They proposed that in deciding whether or not to listen to a message, individuals first rely on cues relating to the message source. If they believe that the critic has deconstructive motives, they choose to reject the criticism and do not process the critic's argument. If, however, they believe that the critic has constructive intentions, they decide to evaluate his or her arguments – if the argument is sound, they accept the criticism. However, if the argument is weak, they reject the criticism. Given that group membership is often used as a cue for inferring motives, with outgroup critics being seen as having destructive motives, outgroup critics may not stand a chance in persuading ingroups even if they have strong arguments. Overall, then, developing sound arguments may not be a sufficient strategy for reducing defensive reactions to outgroup criticism.

Esposo et al. (2013) tested this model in an experiment using 188 Australian university students. Participants read a criticism that was ostensibly made by either an Australian (ingroup) or a non-Australian (outgroup). Moreover, the criticism was accompanied by a strong justification (e.g., government statistics, scholarly citations), weak justification (e.g., personal opinion), or no justification (control condition).

Esposo et al.'s results indicated that there was a main effect of the critic's group membership. Specifically, compared to outgroup critics, ingroup critics elicited higher ratings of constructiveness, agreement, likeability, need for reform, (e.g., believing that there was a need to "act in response to the author's message") and lower ratings of negativity. There was also a significant interaction between critic group membership and argument quality on ratings of agreement, likeability, need for reform, and negativity, but not on ratings of constructiveness. More specifically, when reading criticism from an ingroup member, participants liked the critic more and agreed with the message more when the criticism was supported by strong arguments than when it was supported by weak or no arguments. Participants also evaluated the comments less negatively and indicated a greater need for reform when the criticism was accompanied by strong or no arguments compared to weak arguments. On the other hand, when the criticism came from an outgroup member, argument quality had no effect on ratings of likability, negativity, or need for reform. However, participants were more likely to agree with outgroup critics when they presented strong or weak arguments compared to no argument.

Overall, these findings suggest that argument quality matters to a lesser extent when the criticism is made by an outgroup member than when it is made by an ingroup member. According to Esposo et al. (2013), perceiving the critic's intentions as constructive is an important first step in intergroup persuasion. When no other information is available, individuals may use group membership as a cue to determine the critic's motives. If the critic belongs to an outgroup, he or she may be assumed to have destructive motives and his or her argument may be immediately rejected. On the other hand, if the critic belongs to an ingroup, his or her argument is evaluated on the basis of quality with stronger arguments generally resulting in greater persuasion than weaker arguments. These findings suggest that, although argument quality is important in eliciting attitude change, such a strategy will only be effective if the critic first convinces others that he or she is well-intentioned.

2.3.2 Spotlighting

Spotlighting occurs when critics emphasize that their comments only apply to a specific section of a group. Stating a belief that "Americans are lazy" only applies to Americans who eat at McDonald's is an example of spotlighting. According to Esposo et al. (2013), this strategy is intuitively appealing. We might expect individuals to feel less threatened by criticism directed at a portion of their ingroup because, even though ingroup members may still experience threat to their group identity, they are less likely to experience personal threat. That is, in knowing that the critic does not seek to overgeneralize his or her remarks, individuals may assume that the critic does not intend for his or her comments to relate directly to them. Referring to the previous example, if Mary does not eat at McDonald's, she will not experience threat to her sense of self, resulting in less defensiveness. However, given that Hornsey (2005) argues that the key to reducing defensive reactions is for the critic to appear well-intentioned, spotlighting should only be effective if it simultaneously creates the impression that the critic has the group's best interests at heart.

To investigate this question, Hornsey et al. (2008) conducted an experiment using 82 Australian undergraduate students. Participants read a criticism purportedly written by an Australian (ingroup) or a non-Australian (outgroup). The criticism was either accompanied by spotlighting (e.g., "some of them (us) are fairly racist…of course, not all Australians are like that, but many are") or it was not. Consistent with previous research, there was a main effect of critic group membership such that the ingroup critic was agreed with more, was seen as more constructive and likeable, and elicited fewer negative reactions than the outgroup critic. However, there was no main effect of strategy, indicating that the effects of spotlighting were not significantly different than the effects of not spotlighting. Said differently, directing criticism towards only a portion of a group did not elicit different reactions than directing it at the entire group.

Overall, these results indicate that spotlighting is not an effective strategy for reducing defensive reactions to outgroup criticism. Hornsey et al. (2008) suggest that this strategy is likely ineffective because it does not alter message recipients' attributions of the critic's motives. That is, directing one's criticisms towards only a subset of a group does not necessarily imply that the critic has more benevolent intentions. Indeed, the data indicated that critics who engaged in spotlighting did not significantly differ in ratings of constructiveness than those who did not use spotlighting. Overall, it appears that qualifying criticism does not alter the perception that the comments were motivated by constructive versus destructive reasons and therefore does not reduce defensiveness reactions.

2.3.3 Powerful outgroups

Although individuals may react negatively to criticism, such as by expressing their disagreement, they may still be inclined to change their behaviour. Although most research in social psychology suggests that individuals are inclined to act in accordance with their attitudes, other research suggests that strong situations can make individuals act inconsistently with their attitudes. For example, because job interviews are associated with strong social norms, even if an individual is unenthusiastic about the job, he or she will still express interest in the position in order to maintain a positive impression.

Brander and Hornsey (2006) suggest that powerful others constitute a strong situation. Specifically, they argue that powerful outgroup members may be more effective in eliciting behavioural change in ingroups than less powerful outgroup members. Outgroup members are considered powerful when they are able to punish and reward the ingroup and/or have control over the ingroup's resources (Brander & Hornsey, 2006). Individuals may act in accordance with a powerful other's expectations even when such behaviour betrays their beliefs because doing so enables them to attain rewards and avoid punishment.

Brander and Hornsey (2006) investigated this question within the context of the Iraq war. Eighty-three Australians undergraduate students read an excerpt that criticized Australia's lack on involvement in the Iraq war. The criticism was written by either an Australian (ingroup) or a non-Australian (outgroup). Moreover, the ingroup/outgroup member was either a senior official of the department of foreign affairs (power) or a small business owner (no power). Before and after reading the criticism, participants completed an attitude measure ("To what extent do you feel Australia should support a war with Iraq without UN approval?") and a reverse-coded behavioural measure ("…consider signing a petition against a war in Iraq" and "…voting against a war in Iraq if it were put to a referendum?"). Participants also evaluated the negativity of the comments (e.g., offensive).

The results revealed that there was no main effect of critic profession, indicating that the powerful critic was no more persuasive than the less powerful critic. Moreover, contrary to expectations, there was no interaction between critic profession and nationality. Thus, a powerful outgroup critic was no more effective in eliciting attitude and behavioural change than a less powerful outgroup critic. These results suggest that tasking an influential member of an outgroup with the job of persuading ingroup members may not be an effective strategy.

2.4 Summary

Based on our review of the social identity literature, we recommend the following when trying to persuade members of other groups. Use strategies that clearly demonstrate that you have benevolent intentions. These include 1) attaching praise to criticism, 2) acknowledging your own group's failings on the same criterion for which you are criticizing the other group, and 3) invoking a common identity by referring to a superordinate identity that both you and the other group share (especially one with which the other group members strongly identify) and using inclusive language. When attempting to target the other group's behaviour (and not their attitudes), we recommend making internal attributions about the group's shortcomings.

Strategies to avoid when trying to persuade outgroup members attitudes and behaviours are 1) relying on argument quality. Having a strong argument is necessary when persuading individuals who are systematically processing information. However, before using such a strategy, outgroup critics need to make sure that the other group perceives them as having constructive motives. Other strategies to avoid are 2) using power to get what you want and 3) directing criticism to only a subset of the group (i.e., spotlighting).

This page intentionally left blank.

3. References

3.1 References for Rumours and Social Media

Abad-Santos, A. (2013). Reddit's 'Find Boston Bombers' founder says 'it was a disaster' but 'incredible.' *The Wire.* Retrieved from http://www.thewire.com/national/2013/04/reddit-find-boston-bombers-founder-interview/64455/

Adams, W. L. (2011, August 8). Were Twitter or Blackberrys used to fan flames of London's riots? *Time Magazine Online.* Retrieved from http://content.time.com/time/world/article/0,8599,2087337,00.html

Allport, G. W., & Postman, L. (1947a). *The psychology of rumor.* NY: Henry Holt and Co.

Allport, G. W., & Postman, L. (1947b). An analysis of rumor. *Public Opinion Quarterly, 10*, 501-517.

Baijal, S. (2013, April). Sunil Tripathi: Social media wrongly accuses missing Brown student as Boston Bombing suspect. *Mic.com.* Retrieved from http://mic.com/articles/36615/sunil-tripathi-social-media-wrongly-accuses-missing-brown-student-as-boston-bombing-suspect

Bernardi, D. L., Cheong, P. H., Lundry, C., & Ruston, S. W. (2012). *Narrative landmines. Rumors, Islamist extremism, and the struggle for strategic influence.* New Brunswick, NJ: Rutgers University Press.

Bowcott, O. (2012, June 28). Riots led to 1,400 imprisoned or held on remand, figures show. *The Guardian.* Retrieved from http://www.theguardian.com/uk/2012/jun/28/riots-prison-figures

Brock, T. C., (1968). Implications of commodity theory for value change. In A. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological foundations of attitudes* (pp. 243-276). New York, NY: Academic Press.

Buckner, H. T. (1965). A theory of rumor transmission. *Public Opinion Quarterly, 29*, 54-70.

Caplow, T. (1947). Rumors in war. *Social Forces, 25,* 298-302.

Cesario, J., Grant, H., & Higgins, E. T. (2004). Regulatory fit and persuasion: Transfer from "feeling right." *Journal of Personality and Social Psychology, 86*, 388.

Chakravorty, J., & Jadhav, R. (2011). Three bombs kill at least 21 in India's Mumbai. *Reuters*. Retrieved from http://www.reuters.com/article/2011/07/13/us-india-blast-mumbai-idUSTRE76C2Y420110713

Dawkins, R. (1976). *The selfish gene.* Oxford, UK: Oxford University Press.

Dean, G., Bell, P., & Newman, J. (2012). The dark side of social media: Review of online terrorism. *Pakistan Journal of Criminology, 3*, 103-122.

Denef, S., Bayerl, P. S., & Kaptein, N. (2013, April-May). *Social media and the police – Tweeting practices of British police forces during the August 2011 riots.* Paper presented at the annual meeting of CHI, Paris, France.

DiFonzo, N. (2008). *The watercooler effect.* New York, NY: Avery.

DiFonzo. N., & Bordia, P. (2007). *Rumor psychology: Social and organizational approaches.* Washington, D.C.: American Psychological Association.

Ennals, R., Byler, D., Agosta, J. M., & Rosario, B. (2010). What is disputed on the web? In *Proceedings of the 4th workshop on Information Credibility,* WICOW '10, (pp. 67-74).

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition* (2nd ed.). New York, NY: McGraw-Hill Book Company.

Fogg, B. J., & Tseng, H. (1999, May). *The elements of computer credibility*. Paper presented at the annual CHI conference on Human Factors in Computing Systems, Pittsburgh, PA.

Friedman, R. A. (2006). Violence and mental illness – How strong is the link? *New England Journal of Medicine, 355*(20), 2064-2066.

Friggeri, A., Adamic, L.A., Eckles, D., & Cheng, J. (2014). *Rumor cascades*. Association for the Advancement of Artificial Intelligence.

Fromkin, H. L. (1972). Feelings of interpersonal undistinctiveness: An unpleasant affective state. *Journal of Experimental Research in Personality, 6*, 178-185.

Guardian Interactive Team, Procter, R., Vis, F., & Voss, A. (2011). Behind the rumours: How we built our Twitter riots interactive. Retrieved from http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter

Guerin, B. (2003). Language use as a social strategy: A review and an analytic framework for the social sciences. *Review of General Psychology, 7,* 251-298.

Gupta, A., & Kumaraguru, P. (2011). *Twitter explodes with activity in Mumbai blasts! A lifeline or an unmonitored daemon in the lurking?* (Unpublished manuscript). Indraprastha Institute of Information Technology, Delhi, IN.

Gupta, A., Lamba, H., & Kumaraguru, P. (2013, September*). $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on twitter.* Paper presented at the 8th annual meeting of eCrime Research Summit, San Franciso, CA.

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. *In Proceedings of the 22nd international conference on World Wide Web* (p. 729-736). International World Wide Web Conferences Steering Committee.

Heath, C., Bell, C., & Sternberg, E. (2001). Emotional selection in memes: The case of urban legends. *Journal of Personality and Social Psychology, 81*, 1028-1041.

Heussner, K. M. (2010, January 15). Enough already! 7 Twitter hoaxes and half-truths. *ABCNews.com.* Retrieved from http://abcnews.go.com/Technology/twitter-hoaxes-half-truths/story?id=9565678

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management, 6*, 248-260.

Hughes, M., & Sanchez, R. (2011, August 16). London riots: Met chief Tim Godwin considered shutting off Twitter. *The Telegraph*. Retrieved from http://www.telegraph.co.uk/news/politics/8704239/London-riots-Met-chief-Tim-Godwin-considered-shutting-off-Twitter.html

Infowars.com (2013, April 17). *Navy seals spotted at Boston Marathon wearing suspicious backpacks?* Retrieved from http://www.infowars.com/navy-seals-spotted-at-boston-marathon-wearing-suspicious-backpacks/

Ioffe, J. (2015). After Boris Nemtsov's assassination, 'There are no longer any limits' *The New York Times.* Retrieved from http://www.nytimes.com/2015/02/28/magazine/after-boris-nemtsovs-assassination-there-are-no-longer-any-limits.html?_r=0

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology,* 50(6), 1141-1151.

Janis, I. L. (1982). *Groupthink: Psychological studies of policy decisions and fiascos.* Boston: Houghton Mifflin.

Jenkins, B. M. (2009). Lessons learned from the Mumbai attacks. *Rand Corporation. Retrieved from http://www.rand.org/pubs/testimonies/CT316.html*

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1420-1436.

Joseph, S. (2012). Social Media, Political Change, and Human Rights. *Boston College International and Comparative Law Review, 35,* 145-188. Retrieved from Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk, *Econometrica, 47*, 263-291.

Journal Sentinel. (2014, October). *New trend in wedding proposals: Marathon finish lines*. Retrieved from http://www.jsonline.com/blogs/news/278278661.html

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*, 341-350.

Kapferer, J. N. (1990). Rumor in the stock exchange. *Communications, 52,* 61-84.

Kapferer, J. N. (1992). How rumors are born. *Society, 29*, 53-60.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53*, 59-68.

Kamins, M. A., Folkes, V. S., & Perner, L. (1997). Consumer responses to rumors: Good news, bad news. *Journal of Consumer Psychology, 6,* 165-187.

Kelley, S. (2004). *Rumors in Iraq: A guide to winning hearts and minds* (Unpublished Master's thesis). Naval Postgraduate School, Monterey, CA.

Klausen, J. (2015). Tweeting the jihad: Social media networks of western foreign fighters in Syria and Iraq. *Studies in Conflict & Terrorism, 38*, 1-22.

Kleinman, A. (2013). Boston bombing subreddit get stern warning from moderators (update). *The Huffington Post.* Retrieved from http://www.huffingtonpost.com/2013/04/18/boston-bombing-reddit_n_3110500.html?

Knapp, R. H. (1944). A psychology of rumor. *Public Opinion Quarterly, 8*, 22-37.

Kolbe, A. R., Hutson, R. A., Shannon, H., Trzcinski, E., Miles, B., Levitz, N. … Muggah, R. (2010). Mortality, crime and access to basic needs before and after the Haiti earthquake: A random survey of Port-au-Prince households. Medicine, Conflict and Survival, *26*(4), 281-297.

Kotz, D. (2013, April 24). Injury toll from Marathon bombs reduced to 264. *The Boston Globe.* http://www.bostonglobe.com/lifestyle/health-wellness/2013/04/23/number-injured-marathon-bombing-revised-downward/NRpaz5mmvGquP7KMA6XsIK/story.html

Kumkale, G. T., & Albarracín, D. (2004). The sleeper effect in persuasion: a meta-analytic review. *Psychological bulletin*, *130*(1), 143.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480-498.

Kundani, L. (2013, July). When the tail wags the dog: Danger of crowdsourcing justice. *New America Media*. Retrieved from http://newamericamedia.org/2013/07/when-the-tail-wags-the-dog-dangers-of-crowdsourcing-justice.php

Kumkale, G. T., & Albarracin, D. (2004). The sleeper effect in persuasion: A meta-analysis. *Psychological Bulletin, 108,* 480-498.

Kumkale, G. T., Albarracin, D., & Seignourel, P. J. (2010). The effects of source credibility in the presence or absence of prior attitudes: Implications for the design of persuasive communication campaigns. *Journal of Applied Social Psychology, 40*, 1325-1356.

Leberecht, T. (2010, January 19). Twitter grows up in aftermath of Haiti earthquake. *CNET Magazine.* Retrieved from http://www.cnet.com/news/twitter-grows-up-in-aftermath-of-haiti-earthquake Linders, D. (2012). From e-government to we-government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly, 29*, 446-454.

Lynn, M. (1991). Scarcity effects on desirability: A quantitative review of the commodity theory literature. *Psychology and Marketing, 8*, 43-57.

Maddock, J., Starbird, K., Al-Hassini, H., Sandoval, D. E., Orand, M. (2015, March). *Characterizing online rumoring behavior using multi-dimensional signatures.* Paper presented at the Computer-Supported Cooperative Work and Social Computing, Vancouver, CA.

Madrigal, A. C. (2013). #BostonBombing: The anatomy of a misinformation disaster. *The Atlantic.* Retrieved from http://www.theatlantic.com/technology/archive/2013/04/-bostonbombing-the-anatomy-of-a-misinformation-disaster/275155/

McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A re-examination of the construct and its measurement. *Communication Monographs, 66*, 90-103.

Memmott, M. (2013). More images posted of accused Boston bomber's capture. *NPR*. Retrieved from http://www.npr.org/blogs/thetwo-way/2013/08/28/216416234/more-images-posted-of-accused-boston-bombers-capture

Mendoza, M., Poblete, B., & Castillo, C. (2010, July). Twitter under crisis: Can we trust what we RT? *Ist Workshop on Social Media Analytics (SOMA '10), Washington, D.C.*

Mihanovic, M., Jukic, V., & Milas, M. (1994). Rumours in psychological warfare. *Socijalna Psihijatrija, 22,* 75-82.

Morozov, E. The Net Delusion 81–82 (2011). "LOL" means "laugh out loud." Definition of LOL, Merriam-Webster Online Dictionary, http://www.merriamwebster.com/dictionary/lol (last visited Jan. 6, 2012).

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Oh, O., Agrawal, M., & Rao, H. R. (2010, February). *Analysis of tweets and rumors during the Mumbai terrorist attack of November 2008.* Paper presented at the meeting of the Centre of Excellence for National Security, Sentosa, Singapore.

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly, 27*, 407-426.

Oh, O., Kwon, K. H., & Rao, H. R. (2010, December). *An exploration of social media in extreme events: Rumor theory and twitter during the Haiti Earthquake 2010.* Paper presented at the 31st International Conference on Information Systems, St. Louis, MI.

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies, 65*, 17-37.

Ozturk, P., Li, H., & Sakamoto, Y. (2015, January). *Combating rumor spread on social media: The effectiveness of refutation and warning.* Paper presented at the 48th Hawaii International Conference on System Sciences, Kauai, HI.

Panagiotopoulos, P., Bigdeli, A. Z., & Sam, S. (2012). "5 Days in August" – How London local authorities used Twitter during the 2011 riots. *International Federation for Information Processing,* 102-113.

Panagiotopoulos, P., Bigdeli, A. Z., & Sams, S. (2014). Citizen-government collaboration on social media: The case of Twitter in the 2011 riots in England. *Government Information Quarterly, 31,* 349-357.

Peterson, W. A., & Gist, N. P. (1951). Rumor and public opinion. *American Journal of Sociology, 57,* 159-167.

Petty, R. E. & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change.* New York, NY: Springer-Verlag.

Pierro, A., Mannetti, L., Erb, H. P., Spiegel, S., & Kruglanski, A. W. (2005). Informational length and order of presentation as determinants of persuasion. *Journal of Experimental Social Psychology, 41*, 458-469.

Procter, R., Crump, J., Karstedt, S., Voss, A., & Cantijoch, M. (2013). Reading the riots: What *were* the police doing on Twitter? *Policing & Society, 23*, 413-436.

Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: Methodological innovation for the analysis of big data. *International Journal of Social Research Methodology, 17*, 197-214.

Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011, July). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1589-1599). Association for Computational Linguistics.

Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Patil, S., Flammini, A., & Menczer, F. (2010). Detecting and tracking the spread of astroturf memes in microblog streams. In *Proceeding of the 20th International Conference Companion on World Wide Web,* (pp. 249-252)

Rosnow, R. L., Esposito, J. L., & Gibney, L. (1988). Factors influencing rumor spreading: Replication and extension. *Language & Communication, 8*, 29-42.

Rosnow, R. L. (1991). Inside rumor: A personal journey. *American Psychologist, 46,* 484-496.

Rosnow, R. L. (2001). Rumor and gossip in interpersonal interaction and beyond: A social exchange perspective. In R. M. Kowalski (Ed.), *Behaving badly: Aversive behaviors in interpersonal relationships* (pp. 203-232). Washington, DC: American Psychological Association.

Sabha, L. (2008). HM announces measures to enhance security. *Press Information Bureau, Government of India.* Retrieved from http://pib.nic.in/newsite/erelease.aspx?relid=45446

Seo, E., Mohapatra, P., & Abdelzaher, T. (2012, May). Identifying rumors and their sources in social netowrks. In *SPIE Defense, Security, and Sensing* (pp. ?). International Society for Optics and Photonics.

Shibutani, T. (1966). *Improvised News: A Sociological Study of Rumor*. New York, NY: The Bobbs-Merrill Company Inc.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 Boston marathon bombing.

Stevens, L. E., & Fiske, S. T. (1995). Motivation and cognition in social life: A social survival perspective. *Social Cognition, 13,* 189-214.

Sunstein, C. R. (2009). *On Rumors: How Falsehoods Spread, Why We Believe Them, and What Can Be Done.* New York, NY: Farrar, Straus, & Giroux.

Taylor, M., Wells, G., Howell, G., & Raphael, B. (2012). The role of social media as psychological first aid as a support to community resilience building: A facebook study from 'Cyclone Yasi Update.' *The Australian Journal of Emergency Management, 27*, 20-26.

Tonkin, E., Pfeiffer, H. D., Tourte, G. (2012). Twitter, information sharing and the London riots? *Bulletin of the American Society for Information Science and Technology, 38*, 49-57.

Tripathy, R. M., Bagchi, & Mehta, S. (2013). Towards combating rumors in social networks: Models and metrics. *Intelligent Data Analysis, 17,* 149-175.

Turner, P. A. (1993). *I heard it through the grapevine: Rumor in African-American culture.* Berkeley, CA: University of California Press.

Turner, R. H. (1964). Collective behavior. In R. E. L. Faris (Ed.), *Handbook of modern sociology* (pp. 382-425). Chicago, IL: Rand McNally.

Turner, R. H. (1994). Rumor as intensified information seeking: Earthquake rumors in China and the United States. In R. R. Dynes & K. J. Tierney (Eds.), *Disasters, collective behavior, and social organization* (pp. 244-256). Newark, NY: University of Delaware Press.

Turner, R. H. (1972). *Collective behavior* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Verma, S. K. (2003, February 21). I would rather die than eat beef, says PM. *The Statesman* (India). Retrieved from http://hoovnews.hoovers.com/fp.asp?layout=displaynews&doc_id=NR20030220670.2_9abd001177eee34d

Vis, F. (2012). Twitter as a reporting tool for breaking news. *Digital Journalism, 1*, 27-47.

Weenig, M. W. H., Groenenboom, A. C. M. J., & Wilke, H. A. M. (2001). Bad news transmission as a function of the definitiveness of consequences and the relationship between communicator and recipient. *Journal of Personality and Social Psychology, 80*, 449-461.

Westerman, D., Spence, P. R., & van der Heide, B. (2014). Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication, 19*, 171-183.

Williams, O. (2011, August 8). London riots: Twitter that caused them? *The Huffington Post*. Retrieved from http://www.huffingtonpost.co.uk/2011/08/08/london-riots-twitter-that_n_920791.html

Zuckerman, E. (2008, March 8). The Cute Cat Theory Talk at ETech, *My Heart's in Accra Blog.* Retrieved from www.ethanzuckerman.com/blog/2008/03/08/the-cute-cat-theorytalk-at-etech.

3.2 References for Ingroup Outgroup Influence

Brander, T. V., & Hornsey, M. J. (2006). Intergroup sensitivity and the war in Iraq: A case of attitudes and intentions diverging. *Australian Journal of Psychology, 58,* 166-172.

Esposo, S. R., Hornsey, H. J., & Spoor, J. R. (2013). Shooting the messenger: Outsiders critical of your group are rejected regardless of argument quality. *British Journal of Social Psychological Society, 52,* 386-395.

Halperin, E., & Bar-Tal, D. (2007). The fall of the peace camp in Israel. *Conflict and Communication Online*, *6*, 1-18.

Hornsey, M. J. (2005). Why being right is not enough: Predicting defensiveness in the face of group criticism. *European Review of Social Psychology*, *16*, 301-334.

Hornsey, M. J., De Bruijn, P., Creed, J., Allen, C., Ariyanto, A., & Svensson, A. (2005). Keeping it in-house: How audience affects responses to group criticism. *European Journal of Social Psychology, 35,* 291-312.

Hornsey, M. J., & Esposo, S. (2009). Resistance to group criticism and recommendations for change: Lessons from the intergroup sensitivity effect. *Social and Personality Psychology Compass, 3,* 275-291.

Hornsey, M. J., Grice, T., Jetten, J., Paulsen, N., & Callan, V. (2007). Group-directed criticisms and recommendations for change: Why newcomers arouse more defensiveness than old-timers. *Personality and Social Psychology Bulletin, 33,* 1036-1048.

Hornsey, M. J., Robson, E., Smith, J., Esposo, S., & Sutton, R. M. (2008). Sugaring the pill: Assessing rhetorical strategies designed to minimize defensive reactions to group criticism. *Human Communication Research, 34,* 70-98.

Hornsey, M. J., Trembath, M., & Gunthorpe, S. (2004). ''You can criticize because you care'': Identity attachment, constructiveness, and the intergroup sensitivity effect. *European Journal of Social Psychology, 34,* 499-518.

Iyer, A., Schmader, T., & Lickel, B. (2007). Why individuals protest the perceived transgressions of their country: The role of anger, shame, and guilt. *Personality and Social Psychology Bulletin, 33,* 572-587.

Judd, C. M., Park, B., Yzerbyt, V., Gordijn, E. H., & Muller, D. (2005). Attributions of intergroup bias and outgroup homogeneity to ingroup and outgroup others. *European Journal of Social Psychology, 35*, 677-704.

Oxford. (2015). "Criticism". In *Oxford Dictionaries online*. Retrieved from http://www.oxforddictionaries.com/us/definition/american_english/criticism?q=CRITICISM

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123-205). New York: Academic Press.

Rabinovich, A., & Morton, T. A. (2010). Who says we are bad people? The impact of criticism source and attributional content on responses to group-based criticism. *Personality and Social Psychology Bulletin, 36*, 524-536.

Saguy, T., & Halperin, E. (2014). Exposure to outgroup members criticizing their own group facilitates intergroup openness. *Personality and Social Psychology Bulletin, 40,* 791-802.

Schmader, T., Croft, A., Whitehead, J., & Stone, J. (2013). A peek inside the targets' toolbox: How stigimatized targets deflect discrimination by invoking a common identity. *Basic and Applied Social Psychology, 35,* 141-149.

Tanis, M., & Postmes, T. (2005). A social identity approach to trust: Interpersonal perception, group membership and trusting behaviour. *European Journal of Social Psychology, 35,* 413-424.

Weiner, B. (2001). Responsibility for social transgressions: An attributional analysis. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 331-344). Cambridge, MA: MIT Press.