Understanding how Pragmatic Factors influence Ambiguity in the Production and Comprehension of Health Information

H T Clelland

PhD

2022

Understanding how Pragmatic Factors influence Ambiguity in the Production and Comprehension of Health Information

Harry Thomas Clelland

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the Degree of Doctor of Philosophy

Research undertaken in the Faculty of Health and Life Sciences

September 2022

Thesis Abstract

Language is inherently ambiguous. Words or phrases often share multiple possible meanings, and changes to the circumstances of a conversation can alter how a person chooses to produce (and interpret) language. This is especially important in the context of health communication because talking about human health is sensitive, increasing the potential for ambiguity (and therefore misunderstandings). This thesis examined pragmatic (contextual) factors that influence the production and comprehension of ambiguous language within a health context. To achieve this goal, we applied the principles of Experimental Pragmatics by manipulating the social circumstances surrounding three ambiguous language forms. Specifically, we examined the use and interpretation of (1) Uncertainty Terms (e.g., 'perhaps' or 'maybe'), (2) Quantifiers (e.g., 'some' or '1-in-X') and (3) Generic generalisations (e.g., 'music helps old people'). Across eight pre-registered Open Science experiments (N = 4143), we found that contextual factors significantly affect language production and comprehension. Speakers will be subtly indirect when delivering uncertain threatening news, rather than using explicit uncertainty terms such as 'possibly'. Speakers also use quantifiers such as 'some' or 'many' to manage threat, by downplaying the likelihood they communicate. The '1-in-X' bias and the severity bias cause overestimations of health risk, but they do so independently. Severe outcomes are not considered more likely to occur when they are qualified with a proportional quantifier (e.g., 'a few'). Finally, universal generics (e.g., 'supplement improves function in men') are inviting to writers motivated by appeal, but they risk exaggerating the importance and generality of research findings. The use and interpretation of speech is governed to a large extent by context, so our findings carry implications for effectively communicating health information to the public. The delicate nature of discussing human health offers an opportune window into the psychology of language.

List of Contents

Acknowledgements

First, a huge thank you to my supervisor, and friend, Matthew Haigh. My level of appreciation cannot be summarised in an acknowledgements page; however, I will be forever grateful for the opportunity I was given on this project, and for your unwavering support, guidance, and mentorship throughout. You will surely supervise more PhD students in the future, but you never forget your first!

To my office family, the wonderful (and weird) people of CoCo, thank you for making my PhD experience truly special. I will miss you dearly. I know that you will all miss seeing the back of my head and a Monster can.

Thank you to the extraordinary lab manager duo, Kris and Amy, for the opportunities, career advice, and (most importantly) the laughs. Thank you to my research intern, Izzy. This thesis may have been a page or two shorter without your help. Thank you to my second supervisor, Liz, for your helpful input and general friendliness.

Finally, my upmost thanks and appreciation to my closest friends and family. You all know who you are and should never underestimate how valuable you are to me.

Author's Declaration

I declare that the work contained in this thesis has not been submitted for any other award. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this commentary has been approved. Approval has been sought and granted through the Researcher's submission to Northumbria University's Ethics Online System.

I declare that the Word Count of this Thesis is 65,450 words

Name: Harry T Clelland

Date: 30th September 2022

Chapter 1. Introduction

1.1. Ambiguity

The word ‘ambiguous’ has two definitions. In some cases, it can mean uncertainty or doubtfulness, whereas in other cases it can represent something with multiple possible meanings (Sennet, 2011). Ironically, the word ‘ambiguous’ is therefore by definition, ambiguous. As highlighted by Grice, there is a distinction to be made between what a sentence *says* and what a sentence *means* (Grice, 1985). Considering that words or phrases typically enjoy a multitude of potential meanings, sentences often display ambiguity in varying extremities.

Consider, for example, your boss saying to you “*if you could get that report to me by the end of the day, that would be perfect*”. There is a clear discrepancy here between what is said and what is meant, because whilst your boss has not directly instructed you to complete the report by the end of the day, it is clear that this speech act is a polite request to do so (Pinker et al., 2008). The decision not to make a direct request (e.g., ‘finish the report today’) is particularly intriguing because indirect speech is inefficient and more susceptible to misunderstanding. For instance, your boss may believe that they have given you a hard deadline (today) despite not stating it baldly on record. Similarly, you may believe that finishing the report by the end of the day would be a marker of your merit as a worker, rather than merely a minimum requirement for the day. Despite this, polite requests and indirectness in speech appears to be all-but-universal (Brown & Levinson, 1987), suggesting that it must serve a clear communicative purpose. When unpacking the social and cognitive mechanisms that underpin speech, examples like this demonstrate a fundamental motivation for a person’s choice of language.

Goffman proposed that we as humans possess a sort of public self-image, known as 'Face', that each person has distinct ownership over (Goffman, 1967). Within this there are two aspects of a person's public self-image. Negative face essentially refers to a person's right to remain un-impeded. In other words, they have liberty to act and are free from being imposed upon. Any action that infringes a person's right to freedom-of-action can be considered threatening to their negative face. Comparably, positive face refers to a person's desire to attain a consistent, positive self-image, particularly in the eyes of others. In other words, a person has a positive image of themselves and requires that others share, approve of and appreciate this image. Any action that damages (rather than builds) the social self-image is considered threatening to a person's positive face (Goffman, 1967).

Brown and Levinson (1987) extended Goffman's theory. They define a set of acts that are to be considered threatening to both the speaker's face (i.e., the person performing a speech act) and the hearer's face (i.e., the recipient of the speech act). For instance, a speaker threatens their own negative face when agreeing to dinner plans that they would rather not attend. That is, by agreeing to dinner the speaker is waiving their right to remain unimpeded (negative face), so that they can avoid threatening the hearer's face by turning down the dinner plans (i.e., showing an unwillingness to attend). There are then acts that threaten the face of the hearer (i.e., recipient). Giving a direct order or repeating a reminder threatens the hearer's negative face because they impose upon a person's freedom of action. Disagreeing with an opinion or giving bad news threatens the hearer's positive face because they contradict a person's desire to be viewed positively by others. Importantly, within a collaborative and constructive society it is in everyone's best interest to preserve each other's face (Goffman, 1967), and acts that achieve this goal are known as acts of face-work (Goffman, 1971).

Brown and Levinson's (1978; 1987) Politeness Theory offers a means by which face-work can be achieved. They reason that a rational individual will seek to avoid face-threatening acts (like those outlined above), and in doing so, will devise linguistic strategies for minimising the perceived threat level of an act. Specifically, a person will take '*redressive action*' in the form of politeness. Put simply, when confronted with the necessity of committing a face-threatening act (i.e., when avoiding the act altogether is not a reasonable option), a person will attempt to 'give face' to the recipient in a way that signals clearly that the threat is not intended.

For clarity, consider again your boss who needs a report completed by the end of the day. To soften the impact of the face-threatening act (i.e., imposing on your right to freedom of action), there are multiple possible politeness tactics that your boss can use. They could use '*positive politeness*' which involves 'propping up' the positive face of the hearer. For example, they may choose to preface their request by saying something like '*you're the most reliable person on the team*' or '*you share my passion for this project*', as an indication to you that, at least in some capacity, you and your boss share the same ambition (i.e., a reciprocal working relationship). Alternatively, your boss could use '*negative politeness*', which is intended to redress (set right) a person's negative face. For instance, your boss might explicitly say '*sorry to burden you with this*' to offset the threat of the request, or they could hedge the request by saying '*you may have to complete this today*' to soften the immediacy of the deadline. There is, however, a conflict that arises when engaging in negative politeness, because a person is often torn between wanting to be seen to 'give face', and wanting to go off record (i.e., avoid the act entirely; Brown & Levinson, 1987). The solution to this dilemma lies in what is termed '*conventionalised indirectness*' and functions as a way for the speaker to indirectly go on record. This was the eventual strategy chosen by your boss. Indirectly stating

that "*if you could get that report to me by the end of the day, that would be perfect*" allows them to demonstrate a preservation of negative face, whilst still achieving their objective. That being, almost unmistakably, to make a request.

The communicative intentions of a speaker become ambiguous as a person attempts to interpret politeness-based speech (such as an indirect request). This is primarily because there are multiple possible motivations one could attribute to the nature of that speech. Imagine you have written a book that you might like to publish, and your friend (after reading the book) says 'perhaps I need to read it again, just to be sure it's ready'. If you were to make a literal interpretation of this statement, you would conclude that your friend does not yet know whether the book is ready to be sent to a potential publisher (uncertain). This interpretation is likely to change however, once your friend's interpersonal motivations are considered. One possibility is that they are being tactful, such that they believe the book is not good enough to be published but have chosen to caveat their opinion (i.e., saying they may have to read it again, rather than saying 'I don't think it's ready to be published'). Conversely, another possibility is that they believe the book is, in fact, good enough to be published, but they are being cautious as a precaution (i.e., understating their enthusiasm for the book in case it is immediately rejected by a publisher). It is often the case therefore that speech cannot not taken at face value, particularly when a speaker intends to (or could intend to) be polite (Bonnefon & Villejoubert, 2006, Holtgraves, 2005).

Whilst the use of politeness strategies are deployed in the interest of others' self-esteem, deviations from perfectly efficient speech make language more difficult to process, and this creates confusion about the true meaning of words during an interaction (Bonnefon et al., 2011b). As such, the decision to use politeness as a tactic for managing face reduces the offensive nature of words, but simultaneously puts the accuracy of communication at risk.

1.2. Polite Misunderstandings

Relatively speaking, the potential consequences of politeness-based misunderstandings are somewhat trivial in low stakes situations (Lee & Pinker, 2010). For example, polite indirectness can be used to ask a friend for a favour, or to borrow gardening equipment from a neighbour. In these situations, the threat to both communicators often remains low and the use of politeness as a linguistic strategy can be largely beneficial (Bonnefon et al., 2011b; Pinker et al., 2008). In many cases however, the consequences are far less trivial. Take for instance aeroplane pilots and the communication errors that can cause their plane to crash (Cushing, 1994). A well-known example involves a co-pilot who believes that the aircraft they are about to fly is unsafe due to excess ice on the wings. Rather than going 'on record' to directly state their opinion, the co-pilot instead makes multiple indirect attempts to express their concern without offending their superior (the captain). In this pursuit they say things like "*you see how much ice is on the wings back there?*" and "*let's check those wing tops again, since we've been sitting here a while*", yet despite these attempts the plane attempts to take off and crashes shortly afterwards (see Gladwell, 2008 for a more complete account). This example demonstrates a key issue in understanding linguistic ambiguity. That is, when the stakes are high (i.e., the conversation has social and/or physical consequences), a person is likely to sacrifice accuracy in the interest of being polite, which in turn increases the likelihood of a misunderstanding between two communicators. In other words, as the stakes increase there is greater need for clarity, yet as the stakes increase there is also greater need for politeness (Bonnefon et al., 2011b).

Of course, the need for a politeness strategy depends on the actual information being communicated. Evidence has demonstrated that more elaborate strategies are needed, the more sensitive the information being delivered (Holtgraves, 2005). As such, considering that the

decision to use politeness is influenced by 'how much' of the social situation needs to be managed, it stands to reason that the 'weightiness' of an action (i.e., how face-threatening the speech act is) can be influenced by a number of interpersonal factors.

1.3. Politeness and Context

It is not the case that each individual situation holds the same level of face concern. Brown and Levinson (1987) proposed that the 'weightiness' of face concern in a situation is determined by three broad categories. These comprise the relationship distance between speaker and hearer, the degree of imposition and the hearer's perceived power over the speaker. According to their model, one can determine how much weight (face-threat) is associated with an act by applying the following formula:

$$Wx = D(S, H) + P(H, S) + Rx$$

The formula depicts that the weightiness (Wx) of an act is the sum of relationship distance (D), the hearer's power over the speaker (P) and the strength of the imposition (Rx). For example, there would be a considerable amount of face-threat in asking your boss for a pay rise, because the speaker (the employee) has little power over the hearer (the boss), and the request is quite imposing (i.e., considerable action would be required from the boss in order to meet the request). Likewise, asking a stranger for a donation to your charity carries face-threat through a lack of familiarity between speaker and hearer (large relationship distance). Importantly in this model, the deployment of politeness strategies increase as a function of weightiness. In other words, politeness increases with the degree of face-threat. A burst of experimental work was published in the subsequent years, to test the robustness of this proposition (and the formula that underpins it). This was following the reiteration of Brown

and Levinson's 1987 theory, originally published in 1978 (Brown & Levinson, 1978; Holtgraves & Bonnefon, 2017).

Considering each variable in isolation, there is some evidence in support of the relationship distance (D) variable. Wood and Kroger (1991) observed the expected politeness of various categories of social relationship, and evidenced (in line with Politeness Theory) that increased distance between two interlocutors (communicators) incurs greater politeness. Additionally, Holtgraves and Yang (1992) placed participants into imaginary scenarios in which they were tasked with making a request. Their results also evidence that increased distance is associated with greater politeness. Some research contradicts these findings however. Lambert (1996) reports no relationship between distance and politeness, whilst Baxter (1984) reports the opposite relationship. However, research has sometimes confused relationship distance (which refers to familiarity) and personal affect (which refers to liking; Slugoski & Turnbull, 1988), and more recent work has demonstrated that fleeting emotional states can be a more powerful determinant for the use of politeness than distance (Vergis & Terkourafi, 2015). In this 2015 paper, researchers gave participants hypothetical scenarios and asked them to interpret a speaker's intention to insult their interlocutor. The focus of their analysis was on a greek collocation, *re malaka*, which can be used to insult a person in one sense, but can also be used to show solidarity in another sense. Their results indicated that whilst relationship distance impacted the perceived intention to insult (best friend vs. stranger), there was a stronger influence of the speaker's perceived emotional state. That is, the use of the utterance was considered an insult when the speaker is thought to be experiencing a negative mood as opposed to a neutral mood (p. 331). As such, the distance (D) variable is likely the most complex and multifaceted of the three (Holtgraves, 2005).

Evidence in support of the power (P) variable is far stronger. Blum-Kulka and colleagues (1985) observed acts of request within Israeli society and found that, consistent with politeness theory, increases in the speaker power over the hearer were associated with less politeness. This was later confirmed by a series of experimental studies (see Holtgraves & Yang, 1990, 1992). Leichty and Applegate (1991) placed speakers into an imaginary role and had them carry out persuasive tasks, subsequently coding the messages they produced for face management characteristics. They found that participants used more face-work strategies in situations where they had relatively little power (authority over the hearer). This pattern of results translates to a digital space, as Danescu-Sudhof-Mizil et al. (2013) evidence that politeness is less frequent among Wikipedia editors and Stack Exchange users who are high on the 'reputation scale' of these sites. In practical terms, users who are more 'established' (powerful) on their respective platforms, are less polite to other users. These findings demonstrate that the magnitude of face management required in a social situation depends largely on the specific power dynamics at play.

Lastly, a number of studies have shown that degree of imposition (Rx) is associated with increased politeness (Gonzales et al., 1990; Lambert, 1996; McLaughun et al., 1983). Across two experiments, Okamoto and Robinson (1997) tested the contextual factors that could determine gratitude expressions in English participants. The results of their first study evidenced that the degree of imposition when holding a door open for someone influences participant responses to the gesture, and the results of their second study evidenced that speakers will use more polite expressions when making a large imposition. Contemporary research also evidences the use of greater politeness strategies under high imposition. Holtgraves and Perdew (2016) placed participants into imaginary scenarios, and asked them to communicate the likelihood of a negative event. This event was either less severe (i.e., the

hearer's car needs a new battery) or more severe (i.e., the hearer's car needs a new transmission). It was found that irrespective of how likely the negative outcome was (either 20%, 50% or 80%), participants softened the communicated likelihood of a more severe (imposing) event compared to a less severe event. Specifically, when communicating that the recipient's car needed a new transmission (rather than a new battery), the language used communicated a lower judged probability of the event occurring. This demonstrates how a speaker will attempt to mitigate particularly imposing speech through politeness, as they will use language that portrays more severe events to be less likely. Similar to power (P) and distance (D) then, contextual circumstances that require a speaker to be particularly imposing on the hearer sparks the use of politeness as a face management strategy.

Whilst Brown and Levinson's framework is broadly applicable to most dynamic social situations, some have argued that the model does not account for other variables that should, in theory, impact politeness. Tracy (1990), for example, has called for the model to be adapted to include interpersonal characteristics. As such, the model is not universally considered to be 'all-encompassing' (Holtgraves, 2005). Instead, it is better that the model's three variables (distance, power, imposition) are viewed as an abstract, higher-level set of variables within which smaller, more specific variables can operate. For example, data from Forgas (1999a, 1999b) has shown that subjective mood state can influence politeness. Whilst this is not explicitly accounted for by Brown and Levinson's framework, the specific variable (mood) can be seen to influence a higher order variable (power). That is, a person who is in a sad mood (compared to a happy mood) finds themselves less confident and therefore has less perceived power in the social situation. Of course, the framework would then predict that a person with lower perceived power will be more polite in their communication.

Perhaps the most useful element of the framework then, is that politeness in a wide variety of individual situations can be explained using an almost universally applicable framework (Holtrgaves & Bonnefon, 2017). Importantly though, the empirical support for the model demonstrates a more central point, that situational context can increase levels of ambiguity through politeness. In other words, having established that politeness strategies create ambiguity by differentiating what is *said* and what is *meant* (see 1.1), and that this ambiguity increases the possibility of misunderstandings between speaker and hearer (see 1.2), experimental studies that have evidenced in favour of politeness theory demonstrate that changes to the social situation can increase ambiguity, and can therefore increase the potential for politeness-based misunderstandings. Whilst it can often be harmless (and sometimes advantageous) to adopt politeness strategies within everyday life (Pinker et al., 2008), it is important for research to understand situations in which politeness can be harmful. This often occurs as a consequence of high-stakes situations, because highly face-threatening contexts are prone to produce the most politeness (Bonnefon et al., 2011b). For this reason, researchers have begun to utilise high-stakes situations as a tool for developing a more fine-grained understanding of face-management strategies.

1.4. Health Communication

As an area of linguistics, health communication is concerned with how information about human health is exchanged between people. The area is notably wide in scope. On the one hand, communicating about health can refer to the public at large, such as national health campaigns, advertisements (e.g., supplements) and news media reports of health research. Conversely, information about health can be exchanged between individuals, such as conversations with friends and family, speaking with a pharmacist or direct one-to-one doctor-patient communication.

Considering the latter, when a doctor is communicating with their patient both parties find themselves in a high-stakes environment. Firstly from the perspective of the patient, when seeking medical attention they face the risk of being given bad news. Bad news in the medical literature is information that alters a person's view of the future for the worse (Almond et al., 2009). In other words, it is information that signifies an oncoming dip in quality of life. This could be a temporary burden, like a treatment that will be costly or painful (Quill & Townsend, 1991), or it could likewise be a permanent prognosis, like a chronic loss of function or a symptom that will not improve (VandeKieft, 2001). The patient therefore faces a potentially great deal of face-threat by communicating with their doctor, because both their negative face (autonomy) and positive face (self-image) may be damaged by the exchange.

Similarly though from the perspective of the doctor, the act of communicating about a patient's health is burdensome on a number of counts. First, the consequences of ineffective communication is likely to impact their patient. Research has shown that medical adherence is improved when doctors and patients communicate effectively (Penn et al., 2011; Zolnierek & DiMatteo, 2009). Moreover, poorly delivered bad news can cause unnecessary stress and negative health outcomes for a patient (Sparks et al., 2007; Street et al., 2009), so the medical professional, in particular, is responsible for ensuring an effective communicative process.

Second, a doctor is unable to apply a 'one size fits all' formula to each individual patient. When communicating about serious life-altering diseases (e.g., cancer), some patients prefer full-disclosure about their diagnosis, whereas some patients prefer to be kept in the dark, making it difficult for physicians to ascertain the preferences of each individual patient (Parker et al., 2001). In fact, a review of literature by Ley (1982) found that the proportion of patients who preferred full-disclosure was between 50% and 90% across samples. Such wide variability can cause friction between doctor and patient, with both parties feeling that they

lack key information (i.e., patient under-informed on their diagnosis, doctor under-informed on patient preferences; Levinson et al., 1993). The upshot is that communication techniques must be tailored, rather than the doctor being able to use one, universally effective strategy across the board.

Third, physicians face the difficult but necessary task of communicating diagnostic uncertainty to patients (Gordon et al., 2000; Henry, 2006). These circumstances are particularly threatening for the doctor, because a lack of certainty reflects poorly on their professional competency. A study by Johnson et al. (1988) asked over three-hundred patients to watch pre-recorded videos of doctor-patient interactions. They found that during visits in which uncertainty is communicated, patients are less satisfied whilst displaying decreased confidence and less trust in their physician. A similar pattern of results is observed in more recent studies of first-hand patient visits. Evidence from these papers show that communicating a degree of diagnostic uncertainty causes patients to be less confident in their physician, as well as to perceive their physician as less competent overall (Blanch et al., 2009; Cousin et al., 2013; Politi et al., 2007; Wiseman et al., 2016). It seems to be the case then, that uncertainty is potentially threatening to the positive face of a doctor, because a distinct lack of certainty damages their positive social self-image (i.e., that of a capable medical professional). At minimum, doctors face a communicative 'double-edged sword' when communicating uncertainty, because patients may either appreciate open honesty (yet be no further forward in establishing their prognosis), or may instead react negatively to the doctor's apparent lack of knowledge (Bhise et al., 2018).

From a psycholinguistic perspective however, these high-stakes situations are vitally important, because whilst often uncomfortable for both communicators, doctor-patient communication (and health communication more generally) offers researchers an ideal test-

bed for establishing a more pinpoint understanding of face-management tactics. For example, consider a doctor who believes that their patient may have Alzheimer's disease, and must communicate this to them before recommending further tests. In this situation, the doctor needs to perform a balancing act with the words that they choose. On the one hand, the doctor is obliged to communicate as accurately as possible, which of course would involve being clear about their professional opinion (i.e., an anticipated diagnosis). On the other hand, the doctor must make the news as minimally psychologically harmful to the patient as possible, whilst also preserving their hopes and outlook (or their face). After all, it is possible that the further tests will come back all clear. Achieving both of these objectives at once is challenging because they require directly opposing linguistic strategies. Accurate speech requires minimal deviation from perfectly efficient communication (Pinker et al., 2008), so to be optimally accurate, the doctor could say "*I think you have Alzheimer's disease*". By contrast, protecting face requires purposeful deviation from perfectly efficient communication, so the doctor could be more indirect and say "*I'm really sorry to say, but I think that you may possibly have Alzheimer's disease*". Neither of these statements effectively achieve both objectives. In the former, the information is accurate but seems to make no attempt at preserving the emotional well-being of the recipient. In the latter, the statement is receptive to the recipient's emotions but is less direct, and therefore less accurate than it could otherwise be. It is for this reason that analysing the language of health professionals when mitigating bad news offers an opportune window into the spontaneous face-management strategies used as part of this pursuit.

Studies that have analysed doctors' communication techniques provide a window into the kinds of strategies used to perform the linguistic 'balancing act', described above, when under face-threat. Del Vento et al. (2009) performed a randomly controlled experiment to test the idea that physicians communicate bad news with implicit (rather than explicit) language.

Their results showed that when delivering bad news specifically, experienced physicians would mitigate their words through various forms of implicit language. For example, they would word the news in a way that separated the patient from the disease, or they would downplay the certainty of a diagnosis. Another technique used is negations, such as the physician might say "*your prognosis is not good*" as opposed to "*your prognosis is bad*" (Fraenkel & Schul, 2008). In fact, a number of papers suggest that negations are a focal element of the language used by doctors when diagnosing patients (Chapman et al., 2001; Dalianis & Velupillai, 2010).

Beyond a doctors' word choice, face-management can also be achieved through the amount of words used to deliver the information. Shaw and colleagues (2012) analysed the delivery style of thirty one doctors delivering a standardised 'breaking bad news' procedure (i.e., informing the immediate family of a patient's sudden death). Their results revealed three distinct categories of delivery style, based on the time they took to actually deliver the news. In some cases doctors would take a '*blunt*' approach, which conveyed bad news within the first 30 seconds of the exchange. In some cases they would take a '*forecasting*' approach, which was a more staged delivery of bad news, within the first 2 minutes of the exchange. Most notably though, in some cases doctors would take a '*stalling*' approach, postponing the delivery of bad news beyond 2 minutes, until the recipient would reach their own conclusion about the event without the doctor explicitly stating the news. In other words, the '*stalling*' approach delays delivery of bad news until the person on the receiving end works it out for themselves. These findings are consistent with the tendency for doctors to distance themselves from the news they're delivering (Baile & Beale, 2003), and they also demonstrate that in the communication of health information, deviation from perfectly efficient speech (indirectness) is used spontaneously to manage the threatening nature of a high-stakes situation. Put plainly,

politeness tactics are used in an attempt to remain accurate whilst minimising the psychological threat posed by health information.

It becomes clear based on the clinical evidence presented here, that health-based scenarios are central to furthering our current understanding of politeness theory. A key argument for this is that medical professionals, who are experts in communicating with their patients, still default to polite language when the threat to both parties is high. That is, despite guidelines and training on how best to 'break bad news' (Baile et al., 2000; Buckman, 2005; Girgis & Sanson-Fisher, 1995), doctors will still choose to communicate in other ways (Helft & Petronio, 2007; Mast et al., 2005) that are consistent with politeness-based tactics for managing face. Outside of health-related contexts, politeness increases the potential for ambiguity and misunderstandings (see sections 1.1 and 1.2), so understanding the dynamics of how face is managed within health-related contexts becomes increasingly important, especially given the real world implications attached to this form of communication. In other words, whilst the consequences of politeness tactics can be trivial in everyday life (Pinker et al., 2008), they are not-so-trivial in conversations about human health. For this reason, it is necessary to pursue a deeper understanding of how linguistic face management tactics manifest themselves within high-stakes situations.

1.5. Polite Misunderstandings about Health

Whilst a number of politeness strategies achieve the goal of protecting a patient's face, these strategies can cause ambiguity that could alter their perspective. For instance, consider again the use of negations. Your doctor uttering "*your condition will not deteriorate*" may implicitly indicate to you that they once considered (or maybe expected) your condition to worsen at some point in time. By contrast, this interpretation is far less likely if your doctor were to utter "*your condition will remain stable*", because this does not alert you to the

concept of a decline (Giora et al., 2004). As such, the use of a negation may be considered a marker of the doctor's expectations (Schul, 2011), and may give the impression that the doctor is being insincere, because they once held the opposite prior expectancy (i.e., they did not disclose a negative outlook they held about the patient's health; Beukeboom et al., 2010). In this example then, it can be seen that a linguistic strategy intended to mitigate information in a high-stakes situation (e.g., saying '*you will not die*' vs. '*you will live*') can cause a discrepancy between the speaker's intended meaning and the hearer's perceived meaning.

Stalling the delivery of bad news may also cause ambiguity. As discussed, deviations from efficient speech are harder to process (Bonnefon et al., 2011b), therefore creating confusion over the true meaning of an utterance. If a doctor were to take a '*stalling*' approach they will attempt to distance themselves from the news they are delivering, which is in direct conflict with the patient's interests because they are eager to understand their doctor's genuine expectancies about their conditions (Brown et al., 2011). As a result, the recipient of the news has to work harder, cognitively, to understand the information they are being given. During this process of disentangling the intended message, implied meanings can become emphasised and can impact both a patients' immediate response, as well as their subsequent social actions (e.g., medical adherence intentions; Burgers et al., 2012). In other words, when the recipient of bad news is left to 'read between the lines', this alters their reaction to the news and can alter their intention to adhere to medical advice. This example taken with the former, demonstrates how within the context of health communication the use of politeness strategies to manage face can cause a 'mismatch of meaning' between two communicators.

1.6. Politeness and Context in Health

Just as everyday situations carry varying levels of face-threat, the magnitude of threat within health communication is determined, in large part, by contextual factors. For example, communicating bad news is reported by physicians to be more stressful (difficult to deliver) when the news is unexpected (Baile & Beale, 2003), suggesting that the presuppositions of the recipient can alter how face-threatening an exchange is for the speaker. Considering this within Brown and Levinson's framework, certain health-based scenarios will therefore require more or less politeness, depending on the circumstances. Compared to a situation in which a patient is expecting (or at least anticipating) bad news, the fact that the news is unexpected, adds to the degree of imposition (Rx) being placed on the recipient as a function of the exchange. Put differently, bad news is less imposing to deliver (and therefore requires less mitigation) when the patient is expectant of the news. This example illustrates that specific word choice in each health-based situation is likely to differ based on context.

These differences are not trivial, particularly because there are practical implications associated with seemingly minor differences in word choice. Stavropoulou (2011) attempted to establish the factors that cause non-adherence to medication, and used data from the European Social Survey (ESS) to explore this issue. The results emphasised the importance of the doctor-patient relationship, evidencing that subtle differences in word choice have the potential to influence a patient's perception of their physician, which subsequently influences their long-term adherence to medication. It seems to be the case then, that as a doctor attempts to perform the linguistic 'balancing act' of being truthful and accurate whilst preserving a patient's psychological well-being, the involuntary use of politeness-based speech increases the potential for ambiguity, which in turn alters the real-world outcomes associated with that information exchange.

Critically though, changes to contextual factors under a set of health-based circumstances can alter how language is produced, as well as how language is interpreted, and this fact is especially useful for researchers. With the goal of empirically testing psycholinguistic theory (such as Politeness Theory), observing how changes to circumstance can impact ambiguity in this context, provides researchers with an opportune set of testing grounds for understanding language production and comprehension. In short, testing how context influences ambiguity in health communication can effectively allow us to understand the psychology of language.

Chapter 2. Aims of this Thesis

2.1. Overview

The primary goal of this thesis is to better understand how pragmatic factors (context) can influence ambiguous language when communicating about health. Specifically, we will focus on three particular forms of ambiguous language that are outlined in detail below (see sections 2.3, 2.4 and 2.5). First, we will look at the use and interpretation of *Uncertainty Terms*, such as 'possibly' or 'maybe'. Second, we will look at the use and interpretation of *Quantifiers*, such as 'some', 'many', or '1-in-X'. Finally, we will look at the use and interpretation of universal statements made with *Generic Language*, such as '*Exercise helps men regain bone mass*' or '*Stroke patients recover arm use with virtual reality*'. To achieve this, we will experimentally manipulate the social dynamics surrounding these forms of language, and observe the impact that this has on the potential for misunderstandings between two communicators. Crucially, our aim is to apply these findings to the current empirical understanding of psycholinguistic theory (see 2.6).

2.2. Experimental Pragmatics

The field of experimental pragmatics combines two largely individual academic traditions into one domain of study. The *'experimental'* part of course refers to the pursuit of empirical data via rigorous and controlled studies (i.e., manipulating context). The *'pragmatics'* part refers to the study of determining the intended meaning of a sentence or phrase. Taken together, the combined field can be essentially summarised as the experimental manipulation of context to understand the meaning of language during social interactions (Noveck, 2018a).

Studies across many sub-facets of linguistics have adopted the principles of experimental pragmatics, and evidenced that situational context affects the perceived meaning of words. Take, for example, the existing literature on scalar implicature. Coined by Grice (1989), the term '*implicature*' refers to the suggested, implicit meaning of a sentence or phrase. Scalar implicatures work on the assumption that words (or expressions) can be ranked in terms of their informativeness, for instance the quantifier '*some*' can be considered less informative than '*all*' (Horn, 1992). The idea from theorists is that when a 'weak' term is used in an utterance, the hearer will autonomously generate alternative terms that are stronger in terms of informativeness (Levinson, 2000). For example, consider the statement "*I have read some of the chapters so far*". When reading this statement, one can infer that the use of the weak (or under-informative) term '*some*', is an indication from the speaker that not all of the chapters have been read. That is, the literal statement can be interpreted with the enriched reading "*I have read some but not all of the chapters so far*". The hearer in this scenario makes this interpretation as an implied meaning, by reasoning that the speaker would have no justification for being under-informative, had *all* of the chapters in fact been read (Horn, 1992).

Importantly, empirical studies of these theories of implicature have shown that the extent to which a person makes an inference is not universal. Rather, inferences made are based largely on the characteristics of a situation. A large study of more than 1300 children and adults by Katsos et al. (2016), investigated the perceived truthfulness of statements qualified with '*some*'. Participants across 31 spoken languages were presented with five boxes, within which was a single object (e.g., an apple). They were asked about the truthfulness of a qualified proposition, and the results indicated that statements like '*some of*

the apples are in the boxes' were accepted as true by almost half of the children, but only 16% of the adults. This, along with other research in the area (Noveck, 2001; Noveck & Sperber, 2007), shows that the scalar inferences (e.g., the jump from '*some*' to '*some, but not all*'), will not be made in every case (i.e., when the literal reading will suffice). Instead, pragmatic enrichments are effortful as they are learned and developed with age. If it is the case (as it appears here) that a person will learn over time to reject the literal for an inferred meaning, it must also be the case that a person will be more or less likely to make an inference depending on their given social situation.

One contextual influence that can alter the likelihood of a scalar inference is the presumed motivations of a speaker. In testing this, Bonnefon and colleagues (2009) conducted three experiments. In their first experiment the authors placed participants into hypothetical scenarios, asking them to imagine that they have joined a club (e.g., a poetry club), and that in this club they are being judged on a piece of work that they produce (e.g., a poem). They receive feedback on their work from their fellow club members, which is either positive (face boosting) or negative (face threatening). In the 'face boost' condition participants are told that '*some people loved your poem*', whereas in the 'face-threat' condition participants are told that '*some people hated your poem*'. Of those who were face-boosted, only 12% of participants reasoned that everyone loved their poem, whereas by contrast, of those who were face-threatened 42% of participants reasoned that everyone hated their poem. This suggests that the interpretive jump from the literally spoken '*some*' to the enriched inference '*not all*' is decreased under face-threat.

Following up this finding, the second and third experiments by Bonnefon et al. (2009) were able to explain why scalar implicatures change under face-threat. It was discovered that under these conditions, the word '*some*' is considered to represent a higher proportion. That is,

the amount of people associated with '*some people hated*' is larger than the number of people associated with '*some people loved*', when all else is equal. Furthermore, the difference between these interpretations can be explained by the speaker's presumed motivations. Participants in the face-threatening condition often thought that '*some people hated your poem*' was used by the speaker, even in the knowledge that actually *all* people hated the poem. Whilst this was considered an inaccurate choice of language, it was also considered a kindness, suggesting that a person receiving face-threatening information appears to understand (and perhaps expect) that the person giving them the information is actively trying to manage their face (i.e., soften the bad news).

As it pertains to this thesis, the above example demonstrates a critical point that we will apply to health communication within the current work. Whilst the scope of research that has been completed in the field of experimental pragmatics spans far wider than scalar implicature (see Noveck, 2018b), research in all relevant domains evidence the central idea that context influences the meaning derived from words. Here, we aim to apply the principles of experimental pragmatics to investigate how this plays out in communication about health.

Research on how people make inferences seems to show that part of how inferences are made comes from information stored in memory (Garnham, 2018). However, as described by Garnham, this is unable to account for the fine-grained details of the inferences made by a hearer during conversation. In this respect, experimental pragmatics is particularly useful because, generally speaking, the goal of this discipline is to determine how intention-reading plays a part in how someone makes an inference (Noveck, 2018b). Put another way, when a hearer is trying to figure out the speaker's intentions through the meaning of their words, there are inevitable gaps in their understanding of these intentions. To fill these gaps, the hearer will make inferences (or assumptions) about what the speaker 'really means'. Experimental

pragmatics aids in a more detailed understanding of this inference-making process, because by running experimental studies, we can determine the theories that best explain how meaning is derived from language.

Specifically, we are focusing on forms of language that can be particularly ambiguous (i.e., the hearer has more 'gaps' to fill), and how changes to social dynamics can influence both the production (i.e., generation as the speaker) and comprehension (i.e., interpretation as the receiver) of these ambiguous speech acts. Health communication is the most useful context within which to study pragmatics in this regard, because as described in chapter 1, communication about health requires a balance of tact and accuracy in situations that are often uncertain.

2.3. Uncertainty Terms

Unlike words or phrases that have a concrete reference, such as '*your car keys*' or '*my thesis*', uncertainty terms do not have a solid referent (Holtgraves, 2014). For example, consider the utterance "*it is possible that you will suffer side effects*". The term '*possible*' in this statement refers to the likelihood that you, the hearer, will experience side effects (presumably from a medication or treatment). However, one could not decipher an exact probability from this statement, such that the likelihood of the event occurring could fall anywhere between a very small probability and a very large probability of suffering side effects. In fact, the degree of certainty attached to the word '*possible*' could be anything greater than zero and anything less than one (i.e., full certainty; Bonnefon et al., 2011b), leaving an exceptionally wide range of plausible likelihoods.

Unsurprisingly, uncertain or probabilistic terms like '*possibly*', '*maybe*' or '*perhaps*' are a source of ambiguity, and increase the likelihood of misunderstandings between two

communicators (Juanchich et al., 2019). One primary (but not exclusive) reason for the ambiguity in uncertainty terms is the expectation of politeness. Bonnefon and Villejoubert (2006) compared the perceived probability of '*possible insomnia*' (lower severity outcome) to '*possible deafness*' (higher severity outcome) and discovered that perceived likelihood was higher when the event was more severe. The authors provide an explanation for this finding. The hearer in a typical conversation will reason that a person who uses the term '*possible*' is not in a position to commit to a stronger phrase (e.g., likely). However, when the news being delivered is particularly severe (threatening), the hearer must consider multiple potential motivations for the use of '*possible*'. One of these possibilities is politeness, so rather than genuine uncertainty, the hearer is aware that the speaker may also be motivated to soften the impact of the bad news they are communicating, and will adjust their perception of likelihood accordingly. In other words, the event is considered more likely to occur when severe, because the genuine likelihood is thought to have been softened by the speaker.

The results of this 2006 study demonstrate how the ambiguousness of an uncertainty term can be exacerbated by the contextual dynamics surrounding their use. It is therefore important to develop our understanding of how context influences the production, as well as the comprehension of uncertainty terms within health communication, because under these high-stakes conditions they are likely to cause a discrepancy between the outlook of a doctor and the outlook of their patient (Pighin & Bonnefon, 2011). Whilst Holtgraves (2014) categorises a wide range of word-types under the 'uncertainty terms' label, such as quantifiers (e.g., few), preferential terms (e.g., like) or frequency terms (e.g., often), this thesis will focus only on probabilistic terms like the ones described in this section. This is because they are the most studied in terms of sensitivity to context (see Harris & Corner, 2011; Holtgraves & Perdew, 2016; Juanchich et al., 2019; Juanchich & Sirota, 2013; Sirota & Juanchich, 2015),

and they arguably hold the most potential for ambiguity and misunderstandings (Bhise et al., 2018).

2.4. Quantifiers

Much like probabilistic uncertainty terms, verbal quantifiers (rather than numerical quantifiers) like '*a few*', '*some*' or '*many*' are non-referential because they refer to a proportion that cannot be pin pointed (Holtgraves, 2014). For instance, imagine your doctor saying '*most people see no long-term health impacts after taking this medication*'. The term '*most*' in this statement only offers a certain amount of specificity, because there is a theoretical minimum and maximum proportion of people being referred to (Szabolcsi, 2010). Importantly, the width of plausible proportions represented by a quantifier makes them a source of ambiguity in communication. If a person is told '*some patients do not survive*', the use of '*some*' could legitimately refer to 10% of patients just as much as it could refer to 90% of patients (Degen & Tanenhaus, 2015).

Experimental studies on the use of quantifiers have shown that, much like uncertainty terms, the proportion that they denote changes with context (Bonnefon et al., 2009; Katsos, 2008; Solt, 2016), meaning their use can be more ambiguous under particular circumstances. For example, the politeness account, which is the expectation that a speaker will soften the delivery of bad news, extends to interpretations of the word '*some*' (Feeney & Bonnefon, 2013). Explained another way, the extent to which a person will consider '*some*' to mean '*not all*' will differ under face-threat, as they are expecting the speaker to politely manage the situation. It is certainly the case then, that the interpretation of non-referential terms (e.g., quantifiers) are complicated by changes to the social situation (e.g., how much politeness is required; Holtgraves & Bonnefon, 2017). As such, proportional quantifiers (e.g., some, many) are the first of two types of quantifier that this thesis will evaluate, primarily because they can

contribute to health-based misunderstandings, but also because they are physician's preferred method of communicating risk to patients (e.g., *a few* patients suffer X; Juanchich & Sirota, 2020a).

The second type of quantifier that this thesis will analyse is the '1-in-X' ratio. These are numeric expressions of likelihood (e.g., 1 in 5 chance, 1 in 36 people), and whilst they are not inherently ambiguous (i.e., they pinpoint an exact likelihood, such as 1 in 3 representing 33.3%), they reliably produce a cognitive bias that causes an overestimation of risk (Pighin et al., 2011; Pighin et al., 2015; Sirota et al., 2014; Sirota & Juanchich, 2019). Something that has a 1 in 8 chance of happening, for instance, is considered more likely than something that has a 5 in 40 chance of happening, even though both ratios represent the same raw probability (12.5%). It is therefore important to understand the factors that can influence the interpretation of these ratios, because when they are used to communicate about health, they cause a discrepancy between the likelihood that is intended for the recipient, and the actual likelihood interpreted by the recipient. That being, they believe the chances of the outcome (X) happening are greater than they truly are (Sirota et al., 2018a).

This thesis will explore whether the interpretation of '1-in-X' ratios change with context. A study by Sirota et al. (2018b) analysed the use of these ratios by health professionals, and discovered that they were the most preferred form of communication (over other ratio formats and percentages), but particularly when the health outcome attached to the '1-in-X' ratio is negative. In other words, health professionals prefer these ratios when the information is face-threatening. We reason here that if the severity of an outcome (context) can influence when a '1-in-X' ratio is produced, then by extension, it is reasonable to expect that contextual changes will impact how they are interpreted. Testing this experimentally is important because despite their widespread use, some research has called for the '1-in-X'

ratios to be eliminated from health communication (Zikmund-Fisher, 2011, 2014), so establishing whether situational factors can exaggerate the overestimation of health risk may provide evidence in support of these calls.

2.5. Generic Language

A sentence or phrase is considered to be 'generic' when it expresses a generalisation about a category (Leslie, 2007, 2008). For example, uttering that '*men grow beards*' is a generic statement because one attributes a category (men) with a group level generalisation attached to the group identity (growing beards). One important characteristic of these statements when used in communication, is that they can be considered true, even in the face of obvious exceptions (Krifka et al., 1995). It is rather obvious, for instance, that not every existing man has grown a beard, yet the generic can still be stated truthfully. To illustrate the point, consider the universal non-generic equivalent of this statement, '*all men grow beards*'. The difference between the perceived truthfulness of the two statements is that in the case of the universal non-generic, the existence of a single man who has not grown a beard is enough to falsify the proposition (i.e., it is enough to conclude that '*not all*' men grow beards; Lazaridou-Chatzigoga, 2019).

Generic statements are ambiguous when used in health communication, because they can be used (intentionally or not) to make general claims that are considered more important and more generalisable than can truly be inferred by the evidence supporting the claim (Cimpian et al., 2010; DeJesus et al., 2019). For example, consider a news media article reporting on the findings of health-based research. If the title of the article is generic, such as '*caffeine improves sexual function in women*', this implies that sexual function is improved in '*all*' women. Importantly however, as long as the data being used to make this claim has a minimum of two women who show the described effect, the author of the media article can

state that '*caffeine improves sexual function in women*' truthfully. Recent data shows that generics like these are considered more important (i.e., higher real-world significance) and more generalisable (i.e., applies to more category members) when compared to qualified versions of the same statements (e.g., caffeine improves sexual health in *some* women; DeJesus et al., 2019).

As it pertains to health communication, research has shown that some individuals consider 'Scientists believe…' to mean '*All* scientists believe' (Haigh et al., 2020). Therefore, it is important to understand whether contextual influences can alter the perception of generics, because they are likely to cause a discrepancy between the true generality of findings being reported, and the implied generality of findings being perceived by a recipient. Leslie et al. (2011) evidence that whilst generalisations are accepted as true most of the time, their acceptance relies on the recipients prior knowledge and available information (e.g., '*canadians are right handed*' is rarely accepted as true based on the readily available alternate property, left-handedness). These findings demonstrate in the first instance that the acceptability of a generic is context-dependent. It may also emphasise the importance of understanding generics in health research reports, because when new data is reported, readers cannot effectively draw from readily available knowledge. Put simply, readers cannot falsify a generic statement when it summarises newly discovered information.

2.6. General Aims and Research Questions

This thesis will apply the principles of experimental pragmatics (see 2.2), to understand how situational factors can influence the ambiguity associated with *uncertainty terms* (see 2.3), *quantifiers* (see 2.4) and *generic language* (see 2.5). Specifically, we will analyse this within the context of health communication. For each of these three language types, we will test both language production (i.e., what situational factors can influence a

person to produce ambiguous language) as well as language comprehension (i.e., what situational factors can influence how a person interprets ambiguous language). Holtgraves and Perdew (2016) note that studies of uncertain language (e.g., uncertainty terms, proportional quantifiers) have tended to focus largely on the comprehension of these terms, rather than their production. As a result, in their 2016 paper the authors first tested language production before going on to test language interpretation, which was particularly beneficial in assessing the communicative dyad (i.e., it was more representative of a dynamic conversation one would have in the real world). In the interest of re-affirming this positive step forward, in this thesis we will observe the perspectives of both speaker and hearer when producing and interpreting ambiguous health-based language. As a result, the research questions addressed in this thesis are as follows:

1. How do pragmatic factors influence the production and comprehension of uncertainty terms? (RQ1)
2. How do pragmatic factors influence the production and comprehension of quantifiers? (RQ2)
3. How do pragmatic factors influence the production and comprehension of generic language? (RQ3)

It is anticipated that the eight experiments completed in the pursuit of answering these research questions will improve the current understanding of psycholinguistic theory (such as politeness theory or the gricean perspective). In addition to this, it may be possible for this thesis to generate a set of more targeted, evidence-based recommendations for communicating health risks to patients. Despite the existence of a widely adopted set of general principles for breaking bad news to patients (Baile et al., 2000), only a very small percentage of doctor-

patient communication studies provide actual recommendations on how to formulate the information (Paul et al., 2009). Here, we aim to establish contextual circumstances that can exacerbate the misunderstandings caused by ambiguity. It may therefore be possible to establish a more fine-grained understanding of how to communicate to patients in a way that maximises both accuracy and tact.

Chapter 3. Methodology

3.1. Overview

This thesis presents a series of experiments, reported in chapters 4, 5 and 6. This current chapter will outline a number of commonalities in the approach taken and the methodology chosen when running these experiments. Our project is entirely open science, meaning that all pre-registrations, materials, data and analysis scripts can be accessed openly on the Open Science Framework (see 3.2). All of our data were collected online rather than in-person (see 3.3), and all participants were subject to seriousness checks as an evidence-based approach for assuring the quality of our data (see 3.4). In all of our studies, we created experimental vignettes that we asked our participants to place themselves in (see 3.5). Finally, in addressing our research questions, the studies we have designed are built closely on peer reviewed published papers in the relevant areas, a decision that aligns with the manifesto for 'slow science' (see 3.6).

3.2. Commitment to Open Science

A large point of discussion in psychological science is the 'replication crisis', which refers generally to the inability of researchers to successfully replicate previously discovered effects (Open Science Collaboration, 2015). For example, Klein et al. (2018) collected data from over 15,000 participants across 36 countries, in an attempt to replicate 28 classic psychological effects. It was discovered that in three-quarters (75%) of these replication attempts, the observed effect sizes were smaller than that of the original study. Moreover, only half of the original effects were replicated to statistical significance in the originally observed direction. Worryingly, this pattern of results suggests that newly discovered effects in

psychology are fairly unlikely to replicate. As is discussed in this section, one way to improve this likelihood of replication is to engage in open science practices.

Underpinning this, one important distinction to be made when discussing the benefits of open science, is the difference between 'reproducibility' and 'replicability'. Reproducible research refers to the idea that the same analysis procedure, conducted on the same (original) dataset should produce the same results. Replicable research on the other hand, refers to the idea that re-running the same study using the same materials, but collecting a new dataset, should yield the same pattern of results (Miceli, 2019). This section outlines the steps taken in this thesis (via open research practices) to achieve findings that are both reproducible and replicable. Specifically, open access to data and code maximises reproducibility, whilst pre-registration and open access materials maximises replicability.

Open science (or open research) refers to the "*transparent and accessible knowledge that is shared and developed through collaborative networks*" (Vicente-Saez & Martinez-Fuentes, 2018, p. 434). In practice, it is the idea that the details of how research is designed, captured and assessed should be shared freely, in the interest of openness and connectivity. Research in recent years has highlighted that as the problems being addressed by research become more complex, there is greater need for open practices to drive collaborative and creative solutions (Beck et al., 2022; Ledford, 2015; Van Noorden, 2015). In conducting empirical studies, this can be achieved in a number of key ways. The first is open sharing of research materials, which is essential for allowing research to 'build upon itself' and allow scientists to properly replicate and validate findings (Czarnitzki et al., 2014). In other words, open materials can allow any field of research to cumulatively advance itself and systematically improve knowledge in that respective area.

Similar to this, open sharing of data is also important for the same reasons (i.e., helping with credibility and reproducibility), but some research has indicated that an added benefit of sharing data specifically, is that researchers can achieve greater visibility and research impact (Beck et al., 2019; Hossain et al., 2016; Woefle et al., 2011). More than this though, open data means that the robustness of research findings can be properly scrutinised. This is crucial because there appears to be a growing lack of confidence in past findings among social scientists (Baker, 2016a). As argued by Baker, scientists are less confident about previously established effects (partly) because of unsuccessful replication attempts, and one of the causes attributed to these failures is underspecification of the research process (Baker, 2016b). This is such that researchers have begun to develop strategies for improving transparency in research, such as the 'transparency checklist' which was designed to ensure that a research report is maximally honest about the claims made and the materials that underpin them (Aczel et al., 2019). Without unjustly assuming that a past 'lack of transparency' is ill-intentioned or deliberate, openly sharing data and materials is necessary to ensure that new discoveries are rigorous, reproducible and replicable. That is, different scientists are more likely to come to the same conclusions when data is freely available.

Another important element to achieving the principles of open science is study pre-registration. Contrary to the landscape of psychological research a decade ago (when virtually no researchers submitted pre-registrations for their studies), the number of new pre-registrations in recent years has been reliably accelerating (Simmons et al., 2021). Perhaps most notably, pre-registration is an important defence against '*p-hacking*' (Simonsohn et al., 2014), which refers broadly to the fact that running more than one analysis when testing a hypothesis, increases the likelihood of a type I error (false positive). Simmons and colleagues (2021) outline that if a researcher does not preemptively decide exactly how they will run their

analysis, then their eventual analysis will be '*p-hacked*' to some degree. As such, the sole solution to this issue is for researchers to properly plan their analysis and to commit to its execution. This is done in the form of a pre-registration, which outlines the key details of analysis before any data collection has taken place. The benefit for researchers is that they are able to demonstrate the active steps they have taken against '*p-hacking*' (and so can take credit for a planned analysis), likewise the benefits for the field is enhanced transparency and a decreased type I error rate across the board (Moore, 2016).

In demonstrating our commitment to open science, across all eight experiments reported in this thesis, we completed a full pre-registration detailing (1) our study hypotheses, (2) independent and dependent variables, (3) sample size calculations, (4) inclusion and exclusion criteria, and (5) our precise planned analysis. These pre-registrations were hosted on the Open Science Framework (OSF, Foster & Deardorff, 2017). Additionally, we made sure that for every experiment, our raw data, as well as all relevant materials used to collect the data were freely available on the OSF. This included a 'survey download' (.qsf) file for each study, which allows any external party to import our exact survey into their own machine and replicate precisely our data collection methods. Finally, our analyses were conducted using R, which is a free and open access analysis software (R Core Team, 2021). Alongside our data and materials, we shared all of our coding scripts used to analyse our data, annotated with descriptions of how we are manipulating the data at every stage. More than this, using two R packages 'httr' and 'qualtRics' we coded our scripts to pull raw data directly from the OSF, rather than a typical script which requires the data to be imported from a local machine. Taken together, this means that we have provided readers with freely available analysis scripts that, in order to reproduce our exact analysis procedure, simply requires the reader to download and

run the script in R. Links to the relevant OSF files are provided throughout this thesis, and access to all of this project's study pages can be found via my profile page (osf.io/ebmht/).

3.3. Online (vs. In-Person) Data Collection

It was once the case in research that a very small proportion of published articles reported data that was sampled online (Schleyer & Forrest, 2000). Of course, how the public use the internet in the modern day is very different to how it was used 20 years ago, and the reported issues with online data collection back then can, by-and-large, be considered an artefact of the times. For example, the majority of internet users at this time were highly educated white males, which impacted the representativeness of a sample, and at the same time, there was a concern that many internet users did not have the technical ability to complete surveys (Granello & Wheaton, 2004). Neither of these issues have persisted (in their original extremity) through to present day, however it remains important to consider the potential issues faced when recruiting participant data remotely, because despite an exponentially better 'reach' (suggested by the label 'world-wide web'), psychological research still struggles to collect samples that represent most of the global population (Henrich et al., 2010a; Henrich et al., 2010b).

One primary concern with regards to the validity of survey data generally, is that the demographic characteristics of survey dropouts are not always random. Coste et al. (2013) analysed the response rates of a French national health survey, and discovered a pattern of incomplete and partially complete responses that could often be explained by sociodemographic characteristics, such as lower education level, older age and a foreign background (i.e., non-french). When applied to online surveys, similar issues are likely to impact the data collection process. For instance, Klovning et al. (2009) raised age-related accessibility concerns for surveys distributed online, and Henrich and colleagues (2010a)

demonstrate that modern psychological research routinely samples from populations with predictable sociodemographic characteristics (i.e., western, educated, industrialised, rich and democratic). One issue faced then, is that a typically representative sample can be made un-representative when dealing with partial or incomplete data. This is because those who decline a request to participate in a survey may be different to those who accept, particularly when concerned with the low response rates of the general population (Vaske, 2008).

Another limitation that should be considered is the potential for '*self-selection*' in the sampling process. As highlighted by Duda and Nobile (2010), when surveys are sent out on an open forum (i.e., a website or mailing list) the researcher does not have control over the sample selection, which introduces the possibility of a number of sampling biases (e.g., those who happened to check the website, or those who seek to influence the findings through vested interest). Whilst these potential limitations may provide reason to question the validity of online data collection, here we will outline our specific sampling methods, and argue that the subsequent data that we collected should be considered valid. Following this, in section 3.4 we will outline the steps taken to assure the standards of our online data (e.g., seriousness checks), and argue that this achieves sufficient data quality.

Across eight experiments, we collected online survey data from 4143 English-speaking adults. For the vast majority of participants (n = 3743) the only inclusion criteria (aside from being aged 18 or over) was the ability to speak fluent English. These wide-spanning inclusion standards were necessary because our aim is to better understand linguistic universals in the communication of health information, and so an inclusive sampling strategy was imperative. This, as well as the nature of our studies, meant that a number of common issues with online recruitment were eliminated. For instance, average median completion time per experiment was 5.2 minutes, with no specialised equipment required for completion. Given that both

reduced interest in a study and increased burden when taking part, are often the cause of dropouts in online studies (Galesic, 2006), our short and straightforward experiments encouraged low attrition rates, improving the representativeness of our sample (Pan & Zhan, 2020). Plus, our studies were hosted on Qualtrics survey platform, meaning they were highly accessible (i.e., could be completed on a PC, tablet or mobile; Snow & Mann, 2013).

Our approach to stopping rules across the entire project was largely consistent. For each experiment we set a minimal sample size target, and the data collection phase would conclude at the point at which recruitment had 'ran out of steam' (i.e., close to zero new daily responses). Response collection would not be terminated before the initial count had reached our minimum sample size target plus ~20% (to allow conservatively for the average dropout in psychological research surveys; Hoerger, 2010). There are two exceptions to this rule, however. In Experiment 2 (section 4.1.7) we issued 200 letters to each be rated by a single participant, and as such we terminated data collection once all letters had been rated once (i.e., a final sample of 200). Then in Experiment 8 (section 6.2) our recruitment was carried out on Prolific, limited by a pot of available funds. In this case we exceeded our minimal sample size target but concluded data collection once the funding pot was empty. Nonetheless, the required sample for all eight experiments was achieved irrespective of recruitment method.

Considering that we saw no value in limiting our sample demographics beyond the ability to speak fluent English, all of our participants were opportunity sampled. Recruitment was carried out entirely online, largely via university mailing lists and recruitment websites (e.g., Survey Circle). In some cases, undergraduate psychology students taking part in our experiments could claim course credit for their participation (2.4% of our total sample), and in some cases our participants were recruited via Prolific and were paid for their contribution

(3.4% of our total sample). The remainder of our sample therefore comprised willing and interested volunteers.

There are many advantages of sampling these data online rather than in-person (face-to-face). First, the data is likely to be more accurate because collection is automated, therefore there is no potential for human error in manual data entry (Lefever et al., 2007). Second, data suggests that participants prefer completing online surveys rather than written surveys (Touvier et al., 2010), so respondents are encouraged to be more honest (Ahern, 2005). Third, the anonymity associated with an online response reduces the temptation to provide socially desirable answers (Reips, 2011). Most importantly though, the benefits of increased power with online studies greatly outweigh the methodological limitations. Fraley and Vazire (2014) have outlined that when a study is underpowered, the researcher can run into a number of statistical issues. They are, for example, less likely to detect a true effect in the population, and more likely to have made a Type I error (false positive) when a significant effect is found. When this happens, it is also more likely that the effect size obtained is 'inflated' rather than representative of the true population effect. Despite this, a recent review suggests that researchers in psychology '*have a pathological fear of overpowered studies*' (Brysbaert, 2019, p. 2). That is, despite the risks associated with insufficient power, current sample size practices do not meet the required standards to achieve the recommended 80% power. So whilst there may be some limitations to online data collection (outlined above), the potential to run highly powered studies via large samples is considerably more valuable. One could argue that less researcher oversight may result in data that differs from in-person methods, but it is certainly the case that a sufficiently powered online study will result in more accurate conclusions than an underpowered in-person study. For the reasons described, the decision to run this project entirely online (rather than face-to-face) is justified.

3.4. Quality Assurance

One issue that arises for researchers when collecting data via web-based surveys is that some participants, who may be just browsing the web, can submit a response without providing properly thought-out answers (Reips, 2009). For example, a fellow researcher may be trying to find out about the study's methodology, or an interested visitor may be looking to identify the research question. When participants do not give a survey their full attention like this, it creates a problem for investigators because it reduces statistical power and contributes to 'noisy' data (Oppenheimer et al., 2009). As a strategy for reducing the impact of non-serious responders in a dataset, researchers can opt to use 'seriousness checks'. This involves directly asking participants to indicate the extent to which their response was provided seriously, and it means that the researcher can readily identify participants who were not giving the survey their full attention (Musch & Klauer, 2002).

An important discovery made whilst seriousness checks were becoming known and used, is that a measure of seriousness taken at the beginning of a survey response is the best predictor of subsequent dropout rates (Reips, 2002). This demonstrates the value in establishing serious responses from participants, and researchers in subsequent years sought to further develop these findings, directly investigating data quality improvement as a result of seriousness checks (Aust et al., 2013). In this 2013 paper, Aust and colleagues made a direct comparison between the quality of data provided by serious responders, and the quality of data provided by non-serious responders, in an attempt to establish whether removing non-serious responders improves overall quality of data. They collected the responses of more than 3700 German adults, who at the end of their survey openly told the researchers that either "*I have taken part seriously*" or "*I have just clicked through, please throw my data away*". Across a number of measures, the authors discovered that the data provided by non-serious responders

behaved significantly differently, and resulted in far more varied and unreliable answers. Additionally, it was found that removing non-serious responses from an analysis improves the quality of the data, even after other data quality checks have already been completed (e.g., screening IP addresses, speedy completion etc.). On the back of these findings the authors recommend the standardised adoption of seriousness checks in all online studies (Aust et al., 2013).

The experiments reported in this thesis satisfy the recommendations of researchers (Aust et al., 2013; Reips 2009). The final question at the end of each of our experiments is a simple request, "*please tell us honestly whether you took the task seriously*". Respondents select either "*I took the task seriously, use my data*" or "*I did not take the task seriously, do not use my data*", and those who openly declare a non-serious response were removed from all analysis. For studies that rewarded participants for their time (e.g., Prolific), respondents were reassured that they would still be credited for their response, irrespective of their answer to our seriousness check. Lastly, the decision to place seriousness checks at the end of our surveys (rather than at the beginning) was purposeful. It first reflects the method of Aust et al. (2013) in evidencing data quality improvements, but it also means that if a participant's motivations were to alter during the study (e.g., starting, but not finishing seriously), then only a check at the end of the survey could capture this.

3.5. Experimental Vignette Studies

It is obvious to state that in order to establish a causal relationship, one must run an experiment. Even though this is simple in concept, a problem that faces experimental designs is that a researcher will sacrifice how generalisable, as well as how externally valid their results are in favour of robust internal validity (Scandura & Williams, 2000). As a result, there is a dilemma in designing a study because on the one hand, running an experiment maximises

internal confidence but causes uncertainty about the generalisability of findings, whereas on the other hand, running a non-experimental (observational) study maximises generalisability but cannot be used to properly infer directionality or causality (Aguinis & Bradley, 2014). The use of experimental vignettes can be utilised as a way of (at-least partially) resolving this issue. Atzmüller and Steiner (2010) describe such vignettes as a short and carefully constructed description of a person, object or situation. For this thesis, a vignette refers to the latter (i.e., a description of a situation), that is manipulated across one or more experimental conditions (i.e., context). Practically speaking, participants imagine themselves in a hypothetical scenario, and are asked either to produce or interpret health-based language as a function of the vignette (set of contextual circumstances) they have been assigned to.

In order to demonstrate their utility, consider two examples of how experimental vignettes have been used in the areas of study also addressed in this thesis. First, consider Sirota and Juanchich (2019), who set out to test whether the '1-in-X' bias (discussed in Chapter 5) could impact health-based decisions. To do this, they presented participants with a hypothetical scenario, within which they imagine that they are planning a trip to Kenya. On this trip, participants face the possibility of contracting a disease, and indicate to the researchers whether they would cancel their trip based on this risk. Importantly, the researchers manipulated their vignettes across two experimental conditions. In one condition, the risk of contracting the disease was expressed with a '1-in-X' ratio (e.g., 1 in 7). In the other condition, the same risk was expressed with a 'N-in-X*N' ratio (e.g., 5 in 35). By comparing the proportion of participants who chose to cancel the trip to Kenya between the two ratio formats, Sirota and Juanchcich (2019) were able to conclude (in line with the established cognitive bias) that risk expressed as a '1-in-X' ratio significantly impacts health-based decision making (i.e., the decision to cancel the trip).

Second, consider Holtgraves and Perdew (2016; Experiment 1), who aimed to establish whether the severity of an event would impact a person's choice of probability term (e.g., somewhat unlikely, possible) when communicating the likelihood of that event. The authors asked participants to imagine that they have recently borrowed their dad's car, and that after running into some mechanical problems, they must communicate to their dad that the car needs a repair. Having manipulated the experimental vignettes, participants either had to communicate that the car needed a new battery (low severity), or they had to communicate that the car needed a new transmission (high severity). By keeping the scenario the same, yet changing the severity (face-threat), the authors were able to evidence that a high severity event alters a person's choice of probabilistic term, used when communicating the likelihood of a negative event. Taken together, these examples demonstrate how manipulating pragmatic (contextual) factors, across experimental vignettes that are otherwise identical, offers an effective way to understand the mechanisms that underpin both language production and language comprehension.

Applied to the pursuit of answering our research questions, vignettes offer a solution to the internal vs. external validity argument when deciding whether to run an experiment. Though they should not be considered a replacement for natural environments, vignettes immerse the participant in a situation that serves to 'simulate' reality, offering researchers greater external validity than traditional experimental methods whilst still allowing both directionality and causality to be inferred with appropriate rigour (Aguinis & Bradley, 2014). By extension, it is reasonable to assume that the effectiveness of our vignettes will partially depend on our participants' level of investment (i.e., the extent to which they engage with the imaginary scenario). We took a number of steps to maximise this engagement, such as 'padding out' some of our vignettes with superficial information, preface instructions ensuring

that participants understand what they will do (and to be ready to engage with it), and conducting seriousness checks (see 3.4) to ensure that 'casual' responders are not brought forward for analysis. Lastly, we ensured that when participants are asked to imagine a situation, they are only assigned to a maximum of one vignette. Whilst this is beneficial because it hides the study aims most effectively, it also maximises interest and engagement, which in turn enhances data quality (Arthur et al., 2021; Galesic, 2006; Gibson & Bowling, 2020).

3.6. Slow (Incremental) Research

Advances in technology and ways of working in modern research has meant that the research process can now be completed far faster than was once possible (Ellis, 2020). Whilst this sounds useful on paper, the comparative ease of data acquisition, combined with a 'publish or perish' culture that places unnecessary emphasis on research volume over calibre (Smaldino & McElreath, 2016), has created a 'fast science' initiative that is harmful for research because it leads to corner-cutting and contributes to the replicability crisis (Frith, 2020). By contrast, the concept of 'slow science', first introduced by Alleva (2006), refers to a more calculated and methodical approach to research that produces new findings less frequently, but of a higher quality. In other words, it is the general belief that research should avoid the pitfalls of trying to 'jump too far ahead' whilst simultaneously sacrificing methodological robustness.

One important way of ensuring that research findings are quality controlled and robust, is to conduct replication-extension studies (Bonett, 2012). That is, designing an experiment in such a way that it attempts to replicate a previously established finding, whilst building on methodologies in order to extend it further. The benefit of studies designed in this way is that it supplements the generalisability of existing findings (Bonett, 2012; Ioannidis, 2018), but it

also ensures that designs are grounded in research techniques that have been properly scrutinised (i.e., peer reviewed).

As it applies to the research reported in this thesis, we commit to slow (or incremental) science in the sense that we designed our studies to largely mirror the methodology of an existing, established paper in the area. For example, in answering our three research questions we build upon psycholinguistic findings drawn from non-health contexts, and test (experimentally) whether those findings replicate when applied to a health context. We also extend knowledge, through both methodological improvements (e.g., strengthening a manipulation) and the expansion of experimental design (e.g., operationalising an extra contextual variable). Bonett (2012) argues for replication-extension-style research to be the standard practice in psychology, which we have committed to here in the spirit of open science (see 3.2). This is because research that aims to replicate defends against dishonest (or at-least non-transparent) research practices (see John et al., 2012), and ensures that reported effect sizes are not inflated, as appears to be the case in varying degrees within the psychological sciences (Klein et al., 2018).

In the three chapters that follow (four, five and six), we present our empirical studies designed to answer the three research questions laid out in chapter two (RQ1, RQ2, RQ3).

Chapter 4. How do Pragmatic Factors Influence the Production and Comprehension of Uncertainty Terms?

4.1. Politeness and the Communication of Uncertainty when Breaking Bad News (Exp. 1, 2)

4.1.1. Abstract

Uncertain language can be used to express genuine uncertainty but can also be used to manage face (e.g., by softening bad news). These conflicting motivations can create ambiguity in health communication. In this two-part experiment, participants assumed the position of a health specialist and wrote a letter communicating either a certain or an uncertain medical diagnosis. This was addressed to either a patient (high face-threat) or the patient's family doctor (low face-threat). Letters written under high face-threat contained more words and more dispreferred markers (e.g., sorry, unfortunately) than those written under low face-threat. The number of explicit hedges (e.g., possibly, maybe) did not differ as a function of face threat. Our data demonstrate that participants spontaneously produced dispreferred markers (but not explicit hedges) to manage face using indirectness. Ratings from a second set of participants indicate that this strategy did not affect the perceived meaning or manner of the message.

Keywords: Politeness; Uncertainty; Facework; Communication; Pragmatics; Experimental Pragmatics

4.1.2. Introduction

Much of the information shared in social and professional discourse is uncertain. This uncertainty can be communicated in various ways, with the most direct method being to use explicit uncertainty terms such as *possibly*, *probably* and *likely*. Consider, for example, a patient who is told by their doctor 'you will possibly experience erectile dysfunction as a side-effect of your new medication'; the term 'possibly' in this example could be interpreted in several different ways. First, the patient may take a literal interpretation, assuming it has been used by the doctor as a genuine marker of uncertainty. However, they may adjust this interpretation if the doctor's interpersonal motivations are considered. Because of the sensitive interpersonal context, the patient may interpret the doctor as being tactful or cautious, leading to other possible conclusions. It may be inferred, for example, that the doctor is deliberately understating the likelihood of the negative event to tactfully soften the bad news (e.g., saying possible when they believe the outcome is probable or even certain; Bonnefon & Villejoubert, 2006) or as a precaution against blame if the prognosis is wrong (Juanchich et al., 2012). One or both of these motivations (tact or caution) may lead the doctor to use language that diverges from their genuine level of certainty. For this reason, a patient may not interpret uncertainty terms at face value. The multiple possible motivations for using uncertainty terms means that they are a major source of ambiguity in linguistic communication (Holtgraves & Perdew, 2016).

The reason the patient may make these pragmatic inferences is due to the possibility that the doctor is engaging in face management (Goffman, 1967). A person's 'face' refers to their public self-esteem. People project a sense of positive identity that is always at stake during social interactions. This identity is sacred to the individual, and for this reason, people are motivated to manage and protect not just their own face but also to the face of others

(Holtgraves, 2005). For example, communicating negative or critical information can threaten the face of the recipient, so speakers are motivated to communicate in a way that minimises this threat. An awareness of this in the moment, will cause a speaker to employ a number of linguistic strategies. These are acts of facework (Goffman, 1967).

One way that a speaker can achieve an act of facework is to engage in politeness (Brown & Levinson, 1987). Those required to perform a face-threatening act (e.g., deliver bad news) can select from a variety of options to reduce the harm imposed on the recipient. The speaker could use 'positive politeness' to bolster the self-esteem of the recipient, but when speech is particularly imposing (e.g., a medical diagnosis) it may be more effectively managed using 'negative politeness'. Negative politeness strategies such as indirectness, hedging or being apologetic look to redress (or 'set right') the recipients' autonomy (their right to remain unimpeded). In effect, the speaker can minimise harm to face by 'softening the blow' of bad news via the way they communicate (Brown & Levinson, 1987).

In our example above (informing a patient that they may have a negative side effect), there is a threat to the patient's face because as a sexually active male, a diagnosis of erectile dysfunction threatens their public self-image. The use of an uncertainty term (e.g., 'possibly') could be used as a negative face management strategy to manage the face of both speaker and hearer. The patient's face is spared because the doctor avoids bluntness (rather than being told 'this medication will give you erectile dysfunction'). Additionally, the doctor manages their own face by accounting for every eventuality (i.e., if erectile dysfunction doesn't occur, they have not misinformed the patient).

Uncertainty terms such as 'possibly' or 'maybe' are inherently ambiguous, as they have no solid referent (Holtgraves, 2014). Their meaning must be inferred by considering the multiple possible motivations of the speaker (i.e., are they genuinely uncertain, or are they

engaging in face management?). Disentangling these motivations is taxing on mental resources, as the hearer attempts to read between the lines (Bonnefon et al., 2011).

Most research on uncertain language conducted to date has focused on how uncertainty terms are comprehended in different contexts. One contextual factor that has been shown to affect the comprehension of uncertainty terms is the degree of situational face threat. For example, informing a patient that they will 'possibly suffer deafness soon' is more threatening than informing them that they will 'possibly suffer insomnia soon', as the consequences of the former are more severe. When the severity of the communicated information (and thus the face threat to the hearer) changes, so does the interpretation of the uncertainty term. Bonnefon and Villejoubert (2006) found that more severe outcomes were perceived to be more probable. In other words, the likelihood of possibly developing deafness was perceived as higher than the likelihood of possibly developing insomnia (all else being equal). This indicates that recipients of uncertain information adjust their interpretations when they infer that the speaker is motivated by face management.

Several studies have shown that the comprehension of uncertain language is affected by the presumed motivations of the speaker (Bhise et al., 2018; Bonnefon et al., 2009; Holtgraves, 2014; Lee & Pinker, 2010; Pinker et al., 2008; Stewart et al., 2018). In contrast, there has been less focus on the production of uncertain language by speakers or writers (Holtgraves & Perdew, 2016). In line with politeness theory, the studies that have been conducted demonstrate that when information is face threatening, the speaker (or writer) will adjust the wording of their message to manage the recipient's face (e.g., Burgers et al., 2012; Holtgraves & Perdew, 2016; Holtgraves & Yang, 1992; Juanchich & Sirota, 2013; Lee, 1993; Sirota & Juanchich, 2015). For example, this may be a speaker who is certain choosing uncertain language (e.g., uttering 'your drinking may be a problem' rather than 'your drinking

is a problem') or an uncertain speaker understating their degree of certainty (e.g., uttering 'perhaps you drink too much' rather than 'you probably drink too much').

Of the studies looking into uncertain language production under face threat, most have presented participants with a hypothetical scenario and asked them to choose from a list the probability term they would use to communicate an uncertain outcome. For example, Juanchich and Sirota (2013) asked participants which of eight verbal probability terms (e.g., 'slightly probable', 'evenly probable', 'rather probable') they would use to communicate to an investor the likelihood of their investments losing value. Participants were instructed to be informative or to engage in face management (e.g., avoid blame or soften the bad news). This motivational manipulation influenced the probability terms chosen. When participants were motivated to engage in face management (e.g., to avoid blame or upset), they selected significantly less certain probability terms than when they were instructed to be informative. This suggests that when instructed to purposefully engage in face management, people may understate the certainty of a negative event.

A limitation of the research conducted by Juanchich and Sirota (2013) is that participants were explicitly instructed to engage in face management, so it does not tell us whether people spontaneously choose their words to save face. In a later study, Sirota and Juanchich (2015) again asked participants to select the probability term they would use to communicate negative information, but this time without the instruction to manage face. Participants were then retrospectively asked to indicate the motivation for their choice. In this study 41.6% of participants indicated face management as a reason for their choice (by indicating that tact or caution was their motivation). The participants who selected face management concerns as a motive selected more uncertain probability terms than those who indicated their motivation was simply to be informative. This indicates that at least some

people do spontaneously engage in face management and that one method of achieving this is to understate the certainty (probability) of negative events.

While previous studies typically asked participants to select from a list of specific probability terms, Holtgraves and Perdew (2016; Experiment 2) used a more naturalistic open-ended approach, in which participants were asked to communicate a face threatening opinion in their own words. This allowed them to measure the spontaneous production of uncertainty terms as well as identifying other face management strategies. Importantly, their method also allowed them to look at the communicative dyad (i.e., the speaker and hearer together). The first part of their study focused on the type of language participants produced when communicating in a face threatening context, while the second part examined how these messages were comprehended by a second set of participants.

In the first part of their study, Holtgraves and Perdew (2016; Experiment 2) used hypothetical scenarios to examine how participants communicated opinions that were potentially face threatening to a fictional recipient. For example, one of the scenarios required participants to explain that the reason a friend was failing at university was (possibly) due to their excessive drinking and partying. Within the scenarios, they manipulated the degree of Face threat. This was achieved by keeping the scenario constant, while changing the referent. In practical terms, the high face threat condition involved breaking a negative opinion directly to the friend. In contrast, the low face threat (control) condition involved communicating this negative opinion to a third party, while the friend was not present. Secondly, the certainty of the opinion was manipulated. Participants were asked to imagine that they were either 20%, 50% or 80% certain of their opinion (e.g., 20/50/80% certain that excessive drinking is the reason their friend is failing).

While previous research conducted under restrictive forced-choice conditions has shown that people select less certain probability terms under face threat, Holtgraves and Perdew (2016) sought to identify the specific types of uncertain language that people spontaneously produce under face threat. They planned to objectively examine the production of uncertain language by measuring the use of linguistic features such as probability terms (e.g., probably, possibly), hedges (e.g., could, might) and dispreferred markers that signal impending bad news (e.g., unfortunately…, I am afraid to say..., sadly...). It was predicted that when participants were communicating an utterance in the high face threat condition, they would produce more uncertain language. However, contrary to their expectations the use of these uncertainty terms was infrequent. The authors therefore chose to focus on linguistic indirectness as a substitute, which was measured using a five-point scale. The more a message deviated from a direct assertion (e.g., 'you're partying too much') the more indirect it was judged to be. They found that language production was significantly more indirect in the face threatening condition. A second set of participants then rated the probability (uncertainty) of the messages (e.g., the probability the friend is failing due to excessive drinking). The more indirect the message was, the more uncertain (lower probability) the messages were perceived to be.

The findings of Holtgraves and Perdew (2016) reveal that face threat causes greater use of indirect language, and that this indirect language is perceived as more uncertain (lower probability). However, one of the most surprising findings was the infrequent use of explicitly uncertain language by participants. This meant the authors were unable to reliably examine specific strategies for communicating uncertainty (e.g., the production of hedges and dispreferred markers). They reasoned that the lack of explicitly uncertain language was due to the informal interpersonal nature of their scenarios, which led participants to use indirectness

as a more subtle, implicit method of communicating uncertainty. They contrasted their scenarios with more formal situations (e.g., a doctor communicating a diagnosis to a patient, or a financial advisor communicating risk to a client), in which they reasoned that explicit methods of communicating uncertainty would be more likely. These formal situations are generally high stakes, where the use of tact and the avoidance of ambiguity are equally critical. In the study presented below we build on the work of Holtgraves and Perdew (2016) by asking participants to produce certain or uncertain messages under high or low face threat, but in a much more formal manner. Specifically, participants were asked to put themselves in the position of a medical specialist and to communicate a certain diagnosis or an uncertain diagnosis to either a patient (high face threat) or to a patient's family doctor (low face threat). We anticipated that asking participants to communicate a consequential negative outcome in a formal manner would encourage the use of explicit uncertainty terms to manage face threat. These formal, emotionally sensitive scenarios provide an ideal testbed for examining specifically how different types of uncertain language are used to manage face threat.

4.1.3. The Present Research

This pre-registered study aimed to extend the findings of Holtgraves and Perdew (2016). Our intention in this study was to create vignettes that encouraged greater use of uncertain language, allowing us to examine the various ways in which uncertain language can be used to manage face. To increase the strength of the face threat manipulation we asked participants to communicate high stakes information in a formal manner to a potentially vulnerable recipient. We achieved this by using a formal context known to have a high degree of face threat; breaking bad medical news to patients (Baile at al., 2000; Bonnefon et al., 2011; Burgers et al., 2012; Rodriguez et al., 2007; Shaw et al., 2012; Sirota et al., 2018).

Participants were asked to put themselves in the position of a health specialist (i.e., a specialist in a specific condition rather than a general practitioner) and to write a letter communicating diagnostic information. Face threat was manipulated by instructing participants to write a letter addressing the family doctor of a patient (low face threat condition) or the patient themselves (high face threat condition). Communicating directly with the patient is particularly face threatening because the speaker must manage both the recipient's face and their own face while also bridging perceived differences in power, status and social distance between doctor and patient (Brown & Levinson, 1987). The vignettes were also manipulated so that the diagnosis was either certain or uncertain. Certainty was manipulated because speakers who are motivated to manage face may alter the communicated likelihood of a negative event (Sirota & Juanchich, 2015). That is, the use of uncertain language as a politeness tactic appears, at least in part, to be anchored to the speakers' genuine level of (un)certainty. This manipulation allowed us to investigate whether face management strategies differ as a function of (un)certainty. In part 1 we measured the number of hedging terms used, number of dispreferred markers used, time spent writing the letters, and total number of words used to complete the letters. In part 2, a second set of participants rated the letters that were written by participants in part 1. The letters were rated for subjective politeness, certainty, directness and probability (i.e., probability that the patient has the diagnosed condition).

We predicted that the face threat manipulation would affect the number of hedges and dispreferred markers produced by participants who wrote letters in part 1 and also the perception of these letters in part 2 (see Appendix 1 for our detailed definitions of uncertainty terms and dispreferred markers, osf.io/p8rc2). Specifically, we predicted that participants would produce more hedges and more dispreferred markers when the letter was written in the

high face threat condition than in the low face threat condition, as these are acts of negative politeness that can be used to manage face (Brown & Levinson, 1987). Existing production studies illustrate that hedging, in particular, may take many forms in the pursuit of negative politeness, such as the use of a modal (e.g., I feel I *must* inform you) to alleviate responsibility on the part of the speaker (see Markkanen & Schröder, 2010). We also expected that the production of these linguistic devices would result in a greater number of words being produced and a greater amount of time spent writing the letter. These expectations are based on a wealth of data suggesting that breaking bad news takes longer than delivering good news (Dibble & Levine, 2013). In part 2, we expected that letters written under high face threat (relative to low face threat) would be perceived as more polite, less certain and less direct due to the use of negative politeness strategies.

We predicted that the certainty manipulation would affect the number of hedges and dispreferred markers produced by participants who wrote letters in part 1 and the perception of these letters in part 2. Specifically, we predicted that participants would produce more hedges and more dispreferred markers when the diagnosis was uncertain. This is because hedges inherently communicate uncertainty (e.g., maybe, possibly, might) while dispreferred markers may indicate regret at the lack of certainty (e.g., I'm sorry I can't be more certain). We also expected that the production of uncertain language would result in a greater number of words being produced and a greater amount of time spent writing the letter. In part 2, we expected that uncertain diagnoses would be perceived as more polite (because the uncertain language may be perceived as negative politeness), less certain and less direct.

In addition to the predicted main effects of face threat and uncertainty, we also predicted that these factors would interact. In part 1 we predicted that uncertain diagnoses written under high face threat (see lower right cell in Table 1) would result in the greatest

number of hedges and dispreferred markers, the greatest number of words and take the longest to write. In part 2, we predicted that letters written in this condition would be perceived to be the politest. These interaction effects were predicted as we expected the effects of high face threat and uncertainty to be cumulative (i.e., an uncertain diagnosis may include additional uncertain language when there is also the motivation to manage face).

Pre-registration and Data Availability

This study was pre-registered prior to data collection. The pre-registration, materials, raw data and analysis scripts can be found on the Open Science Framework (OSF) (Clelland & Haigh, 2023; osf.io/zu2an/). There are four known differences between our pre-registration and the study reported here. These differences (and our reasoning for them) are discussed below and outlined in more detail on the OSF project page.

4.1.4. Part 1 Method

Part 1 Design

A 2 x 2 independent groups design was used in Part 1. The first independent variable was Face threat; participants addressed the letter to either the patient's family doctor (low threat) or directly to the patient (high threat). The second independent variable was diagnosis certainty; the diagnosis to be communicated was either certain or uncertain. Participants were randomly assigned to one of four independent conditions (see Table 1).

Table 1

The four experimental conditions of our 2x2 independent groups design.

Addressee	*Certain Diagnosis*	*Uncertain Diagnosis*
Doctor (Low Threat)	Certain/Low Threat *n* = 44	Uncertain/Low Threat *n* = 54
Patient (High Threat)	Certain/High Threat *n* = 49	Uncertain/High Threat *n* = 53

We measured the number of uncertainty terms produced by participants. These terms included linguistic hedges (e.g., maybe, might, possibly, potentially). We also measured the number of dispreferred markers (e.g., well, unfortunately, I am afraid, I regret), the amount of time taken to produce the letter and the total number of words written.

Part 1 Participants

A prospective power analysis was conducted using GPower 3.1. For 80% power to detect a 'medium' sized effect (f = 0.25) with an alpha level of .05, we would require a minimum of 128 participants, or 32 per condition (Faul et al., 2007). This was our minimum sample size target. The anticipated effect size used in the sample size calculation (f = 0.25) was comparable to effect sizes reported following similar manipulations. For example, Holtgraves & Perdew (2016; Experiment 2 Part 1) examined the effect of face threat and probability on the production indirect language (using very similar manipulations similar to ours). The effect size of their face threat manipulation was f = 0.32, and the effect size of their probability manipulation was f = 0.34.

Participants were recruited online via opportunity sampling. The study was advertised online through participation websites, social media and university mailing lists. Undergraduate psychology students could claim course credit for completion. The study was open to those aged 18 or over and fluent in English. A total of 462 participants clicked the link taking them to the study information page. After excluding those who did not complete the study (n = 184), declared a non-serious response (n = 4), or identified themselves as having worked as a health professional (n = 40), a sample of 234 completed responses remained. Following manual inspection of the letters, a further 34 were excluded for the following reasons: The letter did not communicate a diagnosis as instructed (n = 18), the letter contained an incorrect

addressee (e.g., letter was addressed to the patient when the task required a letter addressed to the patient's doctor; n = 14), the participant did not engage with the task (e.g., making a comment about the study rather than writing a letter; n = 2).

A document identifying these letters and their reasons for exclusion can be found on the OSF. The remaining 200 letters were written by English speaking adults (140 female, 59 male, 1 selected 'other'), aged between 18 and 63 (Mean age = 27.23, SD = 10.44). Gender makeup did not differ across conditions X^2 (6, N = 200) = 5.75, p = .45.

Part 1 Materials

Six experimental vignettes were created. Each scenario asked the participant to imagine they were a consultant health professional (i.e., a specialist in a single medical condition). They were given information about a fictional patient, a list of recently reported symptoms, details of recently conducted tests and the diagnosis to be communicated (which was either certain or uncertain). Participants were asked to communicate the diagnosis in a letter to either the patient's family doctor, or directly to the patient. An example scenario, as well as key information can be found in Figure 1. All six scenarios can be found on the OSF.

As depicted in Figure 1, six scenarios were created referring to various life-altering diseases. To aid in operationalising the certainty of diagnoses, the chosen diseases have no definitive test and so require a level of professional opinion. There were four conditions, resulting in 24 vignettes. Each of our participants were randomly allocated to write a letter relating to one of these 24 vignettes. The vignettes were padded out with superficial information about the patient, both as a means of enhancing engagement, but also to disguise our manipulations. Superficial information was counterbalanced across scenarios to offset any potential assumptions or biases associated with such information (e.g., we counterbalanced

characteristics such as patient gender, marital status and exercise levels across the six vignettes, see Table 2).

Figure 1

Example scenario, superscripted digits correspond with variable patient information outlined in the subordinate table. Demographics of the fictional patients were counterbalanced across conditions.

Imagine that you are a specialist at a medical practice. You receive some information about a patient, Mr Wilson[1], from his family doctor Dr (David/Emma)* White.

Mr Wilson is 38 years old[2.] He has a partner[3] and a son aged 6[4]. He has been a digital marketing manager for the past 4 years and earns around £36,000 per year[5]. Some of Mr Wilson's interests include supporting his son's football team once per week, dinner dates with his partner and regular trips to the bar with his friends. The majority of his spare time is allocated to family and friends, so Mr Wilson describes himself as un-active, often choosing convenient food options, but he intends to make lifestyle changes[6].

Mr Wilson has been to see Dr White a number of times in recent weeks reporting forgetfulness, difficulty when making decisions and a sense of disorientation while in familiar places. He has recently undergone a brain scan to find out what is wrong.

You see the results and the brain scan shows signs of brain cell degeneration. You think that Mr Wilson has Alzheimer's Disease. Alzheimer's Disease is known to be difficult to diagnose, (so you are uncertain/but you are certain).

TASK - You are tasked with communicating this information in a letter to (the family doctor of Mr Wilson/Mr Wilson). This letter has been started for you below…

Alzheimer's Disease is an irreversible, progressive brain disorder that slowly destroys memory and thinking skills and, eventually, the ability to carry out the simplest tasks.

Dear Dr White/Mr Wilson,

[Participant writes letter here]

Yours Sincerely,
The Specialist Practice

Table 2

Characteristics such as patient gender, marital status and exercise levels across the six vignettes.

Diagnosis	[1]Name	[2]Age	[3]Relationship	[4]Children	[5]Income	[6]Activity
Pancreatic Cancer	Mrs Davison	48	Married	2	£42,000	Medium
Alzheimer's	Mr Wilson	38	Partner	1	£36,000	Low
Multiple Sclerosis	Miss Richardson	22	Single	0	£25,000	High
Crohn's	Mr Henderson	29	Single	0	£30,000	Medium
Parkinson's	Mr Watson	52	Married	2	£44,000	High
Borderline Personality	Miss Johnson	33	Partner	1	£34,000	Low

* Note that the gender of the doctor always corresponded to the gender of the patient

Part 1 Manipulations

Certainty Manipulation

To operationalise certainty within our vignettes, both an uncertain version and a certain version were created. For example, 'Multiple Sclerosis is known to be difficult to diagnose, so you are uncertain' (uncertain condition) or 'Multiple Sclerosis is known to be difficult to diagnose, but you are certain' (certain condition). Rather than manipulating the numeric certainty of the vignettes as in Holtgraves and Perdew (2016; study 2), we operationalised our scenarios as either wholly certain or wholly uncertain. Our reasoning is as follows. Firstly, patient diagnoses and prognoses can rarely be given exact probabilities on a case-by-case basis (Rodriguez et al., 2007). Moreover, there is limited evidence to suggest that people mentally represent probability at such a fine-grained level, as words do not always map neatly onto numeric expressions of uncertainty (Auger & Roy, 2008). We therefore manipulated the

diagnosis to be either wholly certain or wholly uncertain with no numeric or verbal probability terms used.

Face-Threat Manipulation

To operationalise face threat within our vignettes, both a high face threat version and a low face threat version were created, based on the manipulation used by Holtgraves and Perdew (2016). In the high face threat version, participants were tasked with writing a letter addressed to the patient themselves. In the low face threat version, participants are tasked with writing a letter to the patient's family doctor (known as General Practitioner or GP in the UK). Delivering a medical diagnosis is more face threatening when writing directly to the patient, compared to the patient's GP, as the diagnosis directly threatens the patient's self-image. Doctors have described giving a serious diagnosis as 'dropping a bomb' on the patient (Miyaji, 1993), and breaking bad news in a manner that saves face for the patient is particularly stressful on those who are unfamiliar with the process (Ptacek & Eberhardt, 1996). We therefore expected this manipulation to have a strong effect within our sample who were not health professionals.

Part 1 Measures

Hedges & Dispreferred Markers

The pre-registered plan was to automatically count hedges and dispreferred markers using wordcount software. However, it became apparent that this would not be possible to do in automated fashion. The reason is that producing an exhaustive list of hedges and dispreferred markers to use as criteria for word count software is an unachievable task. Detection of hedges and dispreferred markers is largely context dependent and these linguistic devices are often not isolated to a single word (e.g., *I am afraid to say*…). Crompton (1997)

outlines that hedging is best understood as a product of social forces, arguing that it cannot be limited to and labelled as a closed set of lexical terms. Moreover, Fraser (2010) states that '*an expression is usually only recognized as a hedge when it is used in hedging. Thus, it should not be surprising that there is no grammatical class of hedges, since hedging devices are drawn from every syntactic category*' (p. 23). To ensure that linguistic features were considered in their context we decided to code our letters with the help of an independent coder who applied detailed coding criteria.

Our comprehensive instructions for identifying hedges and dispreferred markers were based on the criteria set out by Holtgraves and Perdew (2016). The coding instructions can be found in Appendix 2 (see osf.io/5grxz/). These criteria were based on the working definitions of hedges and dispreferred markers described in Appendix 1.

Criteria for coding the letters was modified slightly from Holtgraves and Perdew (2016). Initially, our intention was to measure hedges (e.g., may, might, could), dispreferred markers (e.g., well, unfortunately, I am afraid, sadly) and probability terms (e.g., likely, probably) separately. However, considering our working definition of hedges, we decided that probability terms should be categorised as a hedge. Our instructions defined a hedge as 'an item of language which a speaker uses to explicitly qualify his/her lack of commitment to the truth of a proposition' (Crompton, 1997; p. 281). Therefore, within the context of breaking bad news to patients, probability terms neatly fit this definition. Terms that allow the speaker to signal caution or probability rather than full certainty, were categorised as hedges. Total number of hedges were summed for each letter.

Our instructions defined dispreferred markers as items of language that 'reduce the face threat to the receiver by sending out a warning that a threat is forthcoming, or acknowledging in a conciliatory manner that a threat has occurred' (Hamilton et al., 2014; p.

199). Language was categorised as a dispreferred marker when a word or phrase signalled that some kind of bad news was coming, or when language was intended to minimise harm. Total number of dispreferred markers were summed for each letter.

First, the instructions were given to an independent coder, who was blind to the hypotheses. These instructions were supplemented with a collaborative example coding exercise with the first author to ensure consistency and understanding. Initial coding was reviewed by the first author and following this a meeting was held in which discrepancies between coder and first author were discussed. Discrepancies were resolved following discussion, with final judgement decisions made by the coder as the independent party. The coder identified a total of 683 hedges and 140 dispreferred markers. Following review by the first author, a further 195 hedges and 28 dispreferred markers were identified. Collaborative discussion of the review also resulted in the removal of 5 of the originally identified hedges and 1 originally identified dispreferred marker. A complete set of coded letters can be found on the OSF. To test the inter-rater reliability of our coding we recruited a further coder to independently code the letters following the same instructions. Intra-class correlations revealed an 'excellent' level of agreement in the identification of both Hedges (ICC = .852) and Dispreferred Markers (ICC = .925) (Hallgren, 2012).

Writing Time

The vignette describing the patient and the text box in which participants wrote their letter appeared on the same web page (see Figure 1 for an example). Time spent writing the letter was measured by calculating the difference in seconds between the first mouse click on that page (when participants are presumed to have clicked on the free text box to begin writing) and the page submit time (when participants clicked to progress to the next page). This is an estimate of writing time as we cannot be sure that the first time a participant clicked

their mouse was to activate the text entry box, although this is the most likely reason (i.e., there was no other reason to click). As such, the difference between first click and page submit was as a proxy for writing time. Because this measure has limitations, we also present exploratory analysis based on total time spent on the page (i.e., total time taken to read the vignette and write the letter, which makes no assumption about the onset of writing).

Number of Words Written

We summed the total number of words used to compose each letter. This was done using the base R function strsplit. A 'word' was defined in our analysis as a string of characters followed by a space.

Part 1 Procedure

This online study was conducted using Qualtrics (Snow & Mann, 2013). Participants first answered some basic demographic questions. Before being presented with a scenario, participants were shown a prompt screen, asking them to ensure a period of uninterrupted attention. Participants were randomly assigned one scenario only, out of a possible 24 (6 scenarios x 4 conditions). The number of participants allocated to each of the 24 scenarios ranged from 5 to 13. Whilst completing the letter-writing task, participants were provided with standardised openings 'Dear Mr/Mrs/Dr…', and conclusions 'Yours Sincerely, The Specialist Practice'. Letters were restricted to a minimum character length of 50 characters, with no maximum. After writing the letter, participants were asked if they completed the task seriously (Aust et al., 2013), asked if they took a break or experienced any distractions during the task, asked about their status as a health professional and finally asked about their experience of giving and receiving bad news. Mean completion time was 10 minutes 16 seconds. A .qsf file for the Qualtrics survey can be accessed via the OSF.

4.1.5. Part 1 Results

Some of the letters written by participants in Part 1 contained spelling or grammatical errors. Prior to analysis, some letters were manually edited. This process was necessary to aid the coding of uncertainty terms, improve accuracy of the word count in Part 1, and also to aid the comprehension of the letters when presented to a separate set of participants in part 2. Re-formatting was completed with the goal of making as few changes as reasonably possible. We corrected typos and misspellings, and we also improved the clarity of utterances that were not initially clear (see OSF for exact changes). For example, participant 232 wrote 'I am writing to inform you of the condition of your male patient you referred to our clinic and undergone a brain scan' (sic); this was corrected to 'I am writing to inform you of the condition of your male patient, who you referred to our clinic having undergone a brain scan'. Details of manual cleaning and the exact changes made can be found on the OSF. No hedges or dispreferred markers were added or removed from the original letters during the editing. The mean letter word count prior to editing was 96.09 words. After editing the mean word count was and 95.73 words.

A 2x2 (Certainty x Face threat) independent groups ANOVA was conducted on each dependent variable (Number of Hedges contained in each letter, Number of Dispreferred Markers, Time Spent Writing, Number of Words Written)[1]. Cohen's *f* is reported as a measure of effect size. All data analysis was conducted using R version 4.0.2 (R Core Team, 2020) and RStudio version 1.3.959 (RStudio Team, 2020). The raw data and analysis scripts for part 1 are publicly available on the OSF.

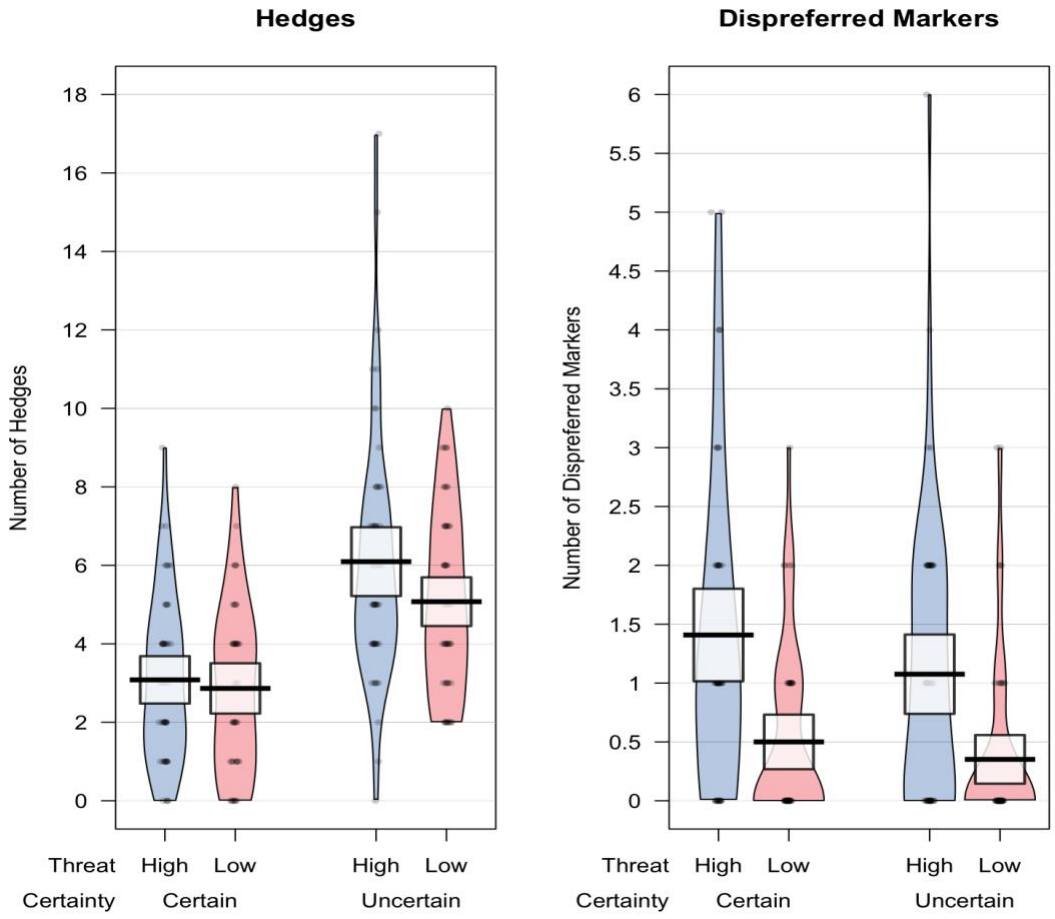
[1] Additional exploratory analysis was also conducted in which Scenario (i.e., the six different vignettes) was also included as a factor (i.e., 2x2x6 Certainty x Face threat x Scenario; see Appendix 3 osf.io/k34tn Scenario did not interact with the other variables (i.e., the effects of Face Threat and Certainty did not differ across the six scenarios).

Hedges and Dispreferred Markers

We measured the total number of hedges and dispreferred markers used in each letter (see Method for details). No further exclusions from the final sample of 200 were made. The mean number of hedges and dispreferred markers in each condition is illustrated in Figure 2.

Figure 2

Mean number of hedges, and mean number of dispreferred markers per letter, across four Certainty x Face-Threat conditions (High/Certain, n = 49, Low/Certain, n = 44, High/Uncertain, n = 53, Low/Uncertain, n = 54). Black horizontal lines represent condition means. Surrounding bands represent 95% confidence intervals. Black jittered dots represent raw data points, and beans represent smoothed density curves.



We found a significant and large main effect of Certainty on the number of linguistic hedges produced per letter $F(1, 196) = 55.196$, $p < .001$, $f = 0.53$. Participants who were asked to communicate an uncertain diagnosis (mean = 5.58), produced significantly more hedges than those who communicated a certain diagnosis (mean = 2.97). There was no significant main effect of Face threat ($F(1, 196) = 3.102$, $p = .080$, $f = 0.10$). For this non-significant effect, the p value was only marginally above 0.05. Participants produced slightly more hedges under high face threat, but the effect size was small. There was no significant interaction between Face threat and Certainty ($F(1, 196) = 1.302$, $p = .255$, $f = 0.07$).

We found a significant and large main effect of Face threat on the number of dispreferred markers produced per letter $F(1, 196) = 29.084$, $p < .001$, $f = 0.38$. Participants asked to communicate a diagnosis directly to the patient (high face threat), produced significantly more dispreferred markers (mean = 1.24) than those who communicated a diagnosis to the patient's family doctor (low face threat; mean = 0.43). There was no significant main effect of Certainty ($F(1, 196) = 2.525$, $p = .114$, $f = 0.11$), and no significant interaction between Face threat and Certainty ($F(1, 196) = 0.372$, $p = .543$, $f = 0.04$).

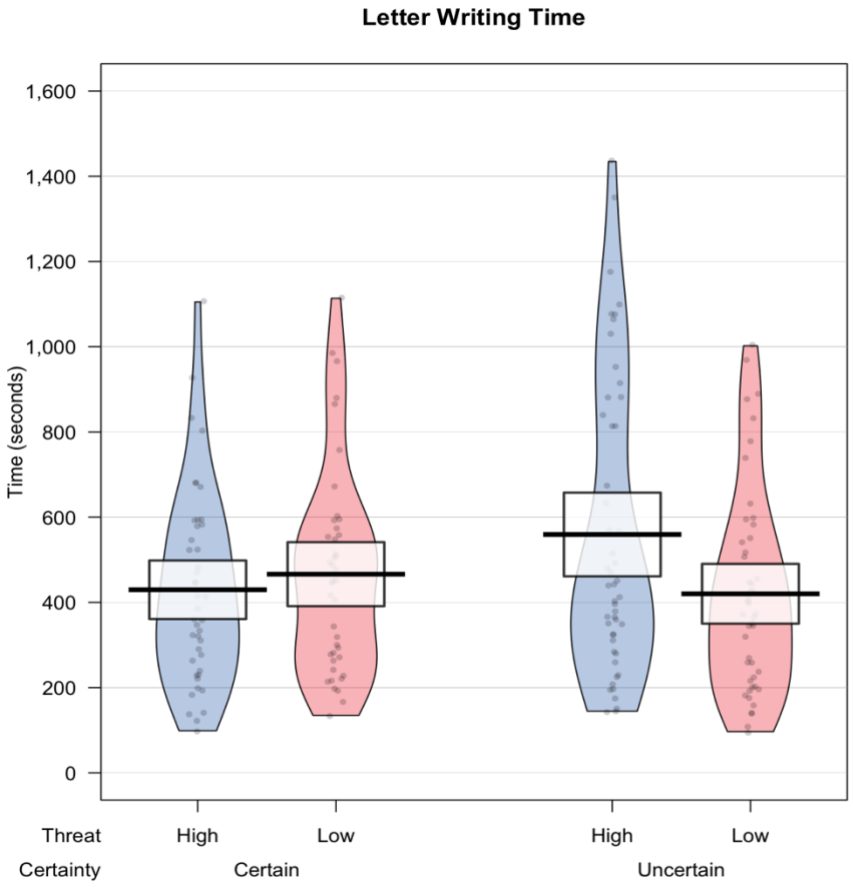
Scenario did not interact with the other variables (i.e., the effects of Face Threat and Certainty did not differ across the six scenarios).

Time Spent Writing (Time from First Click to Page Submit)

For writing time analyses, 14 participants who declared that they were distracted during the task were excluded, leaving a sample of 186 responses. Visual inspection of the data identified possible outliers. We therefore removed any values that were more than 2.5 standard deviations from the grand mean. This resulted in the removal of 3 responses, leaving a final sample of 183. Mean writing time in each condition is plotted in Figure 3.

Figure 3

Mean number of seconds spent writing the letters (this proxy for writing time is the difference in seconds between first click and page submit), as a function of four Certainty x Face-Threat conditions (High/Certain, n = 49, Low/Certain, n = 42, High/Uncertain, n = 50, Low/Uncertain, n = 47). Black horizontal lines represent condition means. Surrounding bands represent 95% confidence intervals. Black jittered dots represent raw data points, and beans represent smoothed density curves.



We found no significant main effect of Face threat ($F(1, 179) = 1.647$, $p = .201$, $f = 0.10$), and no significant main effect of Certainty ($F(1, 179) = 1.094$, $p = .297$, $f = 0.08$) on time taken to write the letter. However, we did find a significant interaction between Face threat and Certainty ($F(1, 179) = 4.825$, $p = .029$, $f = 0.16$). Bonferroni corrected post-hoc comparisons (using $\alpha = 0.025$) revealed that when communicating a diagnosis that is certain,

there is no significant effect of Face threat on writing time $t(82.97) = -0.724$, $p = .471$. However, when communicating a diagnosis that is uncertain, there was a significant effect of Face threat on writing time $t(87.41) = 2.320$, $p = .023$. Those who communicated an uncertain diagnosis under high face threat, took significantly longer to write their letter (mean = 559.4) than those who communicated an uncertain diagnosis under low face threat (mean = 420.2). The pattern of results described above did not differ when the analysis was re-run with outliers included.

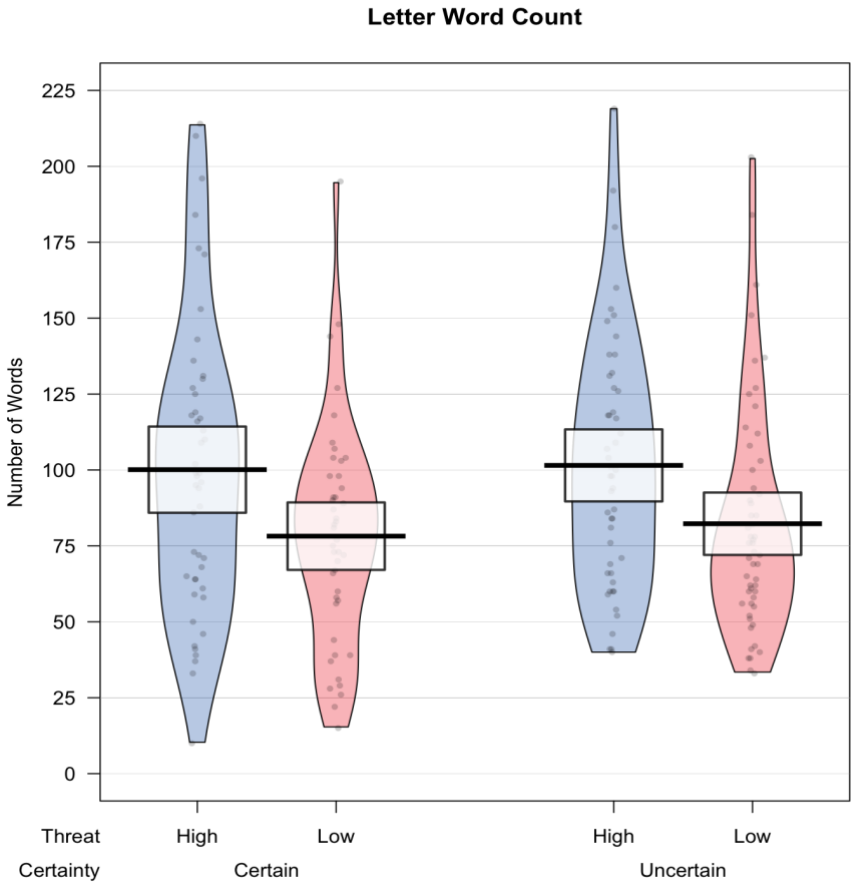
Because our measure of writing time only serves as a proxy for writing time, we also conducted exploratory analysis on the 'page submit' time (see section 4.5.2 for details). This analysis does not make any assumptions about when participants started writing their letters. This analysis produced the same pattern of results as our writing time proxy measure, including a significant interaction between Face threat and Certainty. Details of this additional analysis can be found on the OSF.

Number of Words Written

Word count was calculated by totalling the number of words used to complete each letter. As in our writing time analysis, we excluded participants with word counts that were more than 2.5 standard deviations from the grand mean. Six participants were excluded on this criterion, leaving a final sample of 194. The mean number of words written is illustrated in Figure 4.

Figure 4

Mean letter word count, as a function of four Certainty x Face-Threat conditions (High/Certain, n = 47, Low/Certain, n = 44, High/Uncertain, n = 49, Low/Uncertain, n = 54). Black horizontal lines represent condition means. Surrounding bands represent 95% confidence intervals. Black jittered dots represent raw data points, and beans represent smoothed density curves.



We found a significant main effect of Face threat on the number of words used to produce the letter $F(1, 190) = 12.027$, $p < .001$, $f = 0.25$. Participants who communicated a diagnosis directly to the patient (high face threat mean = 100.8 words) wrote longer letters than those who communicated a diagnosis to the patient's family doctor (low face threat mean = 80.3 words). There was no significant main effect of Certainty ($F(1, 190) = 0.215$, $p = .644$,

$f = 0.03$), and no significant interaction between Face threat and Certainty ($F(1, 190) = 0.051$, $p = .821$, $f = 0.02$). The pattern of results described above did not differ when we re-ran the analysis with outliers included.

4.1.6. Part 1 Discussion

As predicted, diagnosing a serious medical condition under high face threat caused participants to produce more dispreferred markers (e.g., I'm sorry, unfortunately, this must be tough for you), compared to diagnosing a serious condition under low face threat. In other words, when communicating threatening information directly to the patient in our scenarios, rather than to the patient's family doctor, participants produced more terms that signalled and acknowledged the sensitivity of the information being communicated.

Also in line with our predictions, higher face threat caused participants to produce more words when composing their letters than those who wrote under low face threat. When delivering threatening information directly to the patient, participants wrote letters that were longer. This likely reflects the elevated use of dispreferred markers when writing to the patient, which are motivated by face management concerns.

Contrary to our expectations, increasing face threat did not affect the number of hedges (e.g., possibly, likely, potentially) produced by participants. Descriptive statistics did indicate that slightly more hedges were produced under high face threat, but the effect was small and marginally non-significant. The only factor to affect the production of hedges was the writer's degree of certainty. This trivial finding indicates that less certain information was communicated using less certain language. Certainty also had no effect on the production of dispreferred markers or word count. Participants did not routinely use dispreferred markers to indicate regret about their uncertainty (e.g., I'm sorry I can't be more certain).

Perhaps most interestingly, face threat moderated the effect of certainty when we analysed the time taken to write the letters. As predicted, participants took longer to write their letter of diagnosis when addressing it directly to the patient, rather than to the patient's GP, only when the diagnosis was uncertain. Having to manage both sensitivity and uncertainty while communicating a threatening diagnosis, meant that participants required more time to process and engineer their speech in such a way that attends to the sensitivity of the situation, but also accurately depicts their level of commitment to the information being exchanged. This suggests that joint pressures of genuinely communicating uncertain information in a tactful way increased the demands of the task. It is possible that the need to use uncertainty terms to communicate genuine uncertainty required speakers to think of alternative methods to manage face. This interaction effect was observed on our most sensitive measure (writing time) but not observed on our other measures (hedges, dispreferred markers or word count), suggesting that the joint demands of managing face and communicating uncertainty affect processing time but not the language produced.

4.1.7. Part 2

In part 2 of this study we sought to evaluate the communicative dyad, in a way that builds upon the findings of Holtgraves and Perdew (2016). We wanted to understand how the language produced under varying levels of face threat would be interpreted from the receiver's perspective. To measure this, we gave the letters that participants wrote in part 1, to a new set of participants who rated the letters on four interpretive measures.

First, we measured probability (e.g., 'in your opinion, what is the probability that Mr Wilson has Alzheimer's Disease?'). Holtgraves and Perdew (2016) found that probability ratings did not differ between messages produced under high face threat and those produced under low face threat (i.e., participants judged the likelihood of a negative opinion as the same

irrespective of whether it was written in a socially threatening situation or not). However, Bonnefon and Villejoubert (2006) reported evidence that hedging terms such as 'possibly' were perceived to have higher probability in face threatening contexts. Because our face threat manipulation was more formal than that used by Holtgraves & Perdew (2016) and the information to be communicated was more severe, we expected that the linguistic strategies used to manage face threat in part 1 would be more detectable by participants in part 2, leading to higher probability estimates.

Second, letters were rated on certainty (i.e., 'how certain is the specialist about their diagnosis?'). Because uncertainty terms can be used to manage face, we predicted that letters written under high face threat would be perceived as less certain.

Third, letters were rated on directness (i.e., 'how direct is this diagnosis?'). This question was followed by our definition of directness ("*direct language can be defined as blunt and straight to the point*"). Holtgraves and Perdew report that when their scenario was face threatening, participants produced more indirect language, or language that differs from the most efficient way of communicating. We expected therefore, that a diagnosis produced under face threatening conditions would be interpreted as more indirect (i.e., less direct) than a diagnosis produced under less face threatening circumstances.

Fourth, letters were rated on politeness (i.e., 'how polite is this letter?'). Bonnefon and Villejoubert (2006) evidenced that when communicating negative health information to a patient, the use of hedging terms (e.g., 'you will possibly suffer from…'), caused participants to interpret the use of this term as a technique for managing face (i.e., masking the seriousness of the information). We expected that letters written under high face threat would be interpreted as more polite than letters produced under low face threat.

Part 2 Design

There were no further experimental manipulations in part 2. Letters written in part 1 under low or high face threat and communicating a certain or uncertain diagnosis were subjectively rated on the four measures detailed above.

Part 2 Participants

We set out to have each individual letter produced in Part 1 rated by a single participant in Part 2, resulting in a target sample of 200. The study was open to those who were aged 18 or over, fluent in English and had not taken part in Part 1. Each participant was randomly assigned to rate one of the letters generated in Part 1. Following the same recruitment strategies, a total of 219 participants consented to take part. Nine respondents were shown a unique letter, but did not then go on to rate the letter, meaning that our randomiser presented 9 letters twice. Duplicate ratings were removed, and a further 10 incomplete responses were also excluded. A final sample of 200 English-speaking adults (39 male, 169 female, 1 prefer not to say), aged between 18 and 77 (mean age = 30.31, SD = 12.75), rated a single letter. Given that our data from part 1 served as the stimuli for part 2, the number of participants assigned to each condition was the same. Gender makeup did not differ across conditions X^2 (6, N = 200) = 3.33 p = .77.

Part 2 Materials

Each of the 200 letters written in Part 1 were converted into separate experimental vignettes, see Figure 5 for an example. Formatting examples from all 6 scenarios can be found on the OSF.

Figure 5

Example scenario as it appeared to participants in part 2.

Mr Wilson has recently visited his family doctor, Dr David White. He was suffering from forgetfulness, difficulty when making decisions and a sense of disorientation while in familiar places. Dr White conducted a brain scan, and the results of the scan were analysed by a specialist at the medical practice.

Two weeks later, Dr White/Mr Wilson receives a letter that contains a diagnosis. The letter reads as follows…

Dear Dr White/Mr Wilson

Letter written under this set of experimental conditions in part 1 was inserted here

Yours Sincerely,
The Specialist Practice

Part 2 Measures

The four measures described above (Probability, Certainty, Directness, Politeness) were each rated on a scale ranging from 0 to 100.

Part 2 Procedure

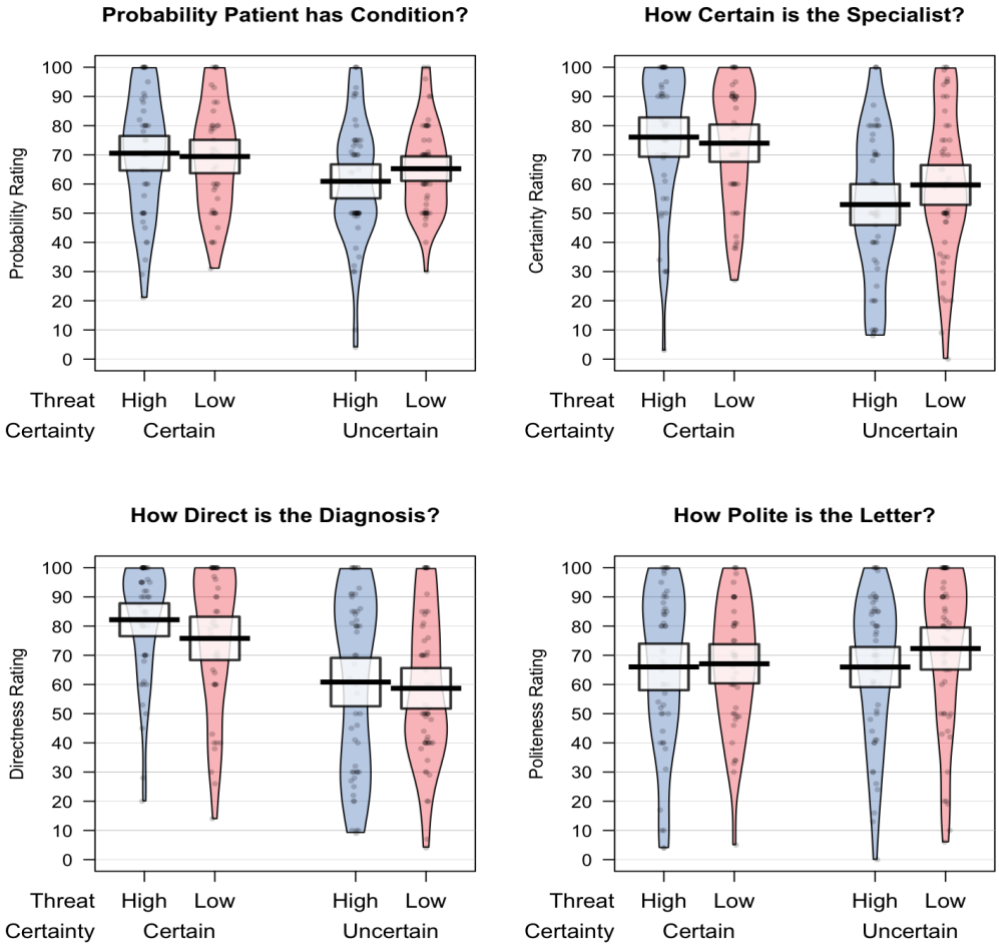
Participants accessed part 2 via an anonymous link. Volunteers were informed that they should not take part if they took part in part 1. They first answered some basic demographic questions. Afterwards, participants were randomly presented with one vignette, containing a single letter written from part 1. Viewing a single letter aids the believability of the hypothetical scenario and avoids revealing the experimental manipulations that were made in part 1. Once participants had read and considered the scenario, they reported their rating of probability, certainty, directness and politeness. We again ran seriousness checks before participants submitted their survey. Mean completion time was 2 minutes 49 seconds. A .qsf file for the Qualtrics survey can be accessed via the OSF.

4.1.8. Part 2 Results

The raw data and analysis scripts for part 2 are publicly available on the OSF. The data from our four measures are plotted in Figure 6.

Figure 6

Mean probability, certainty, directness and politeness ratings across our four experimental conditions (High/Certain, n = 49, Low/Certain, n = 44, High/Uncertain, n = 53, Low/Uncertain, n = 54). Black horizontal lines represent condition means. Surrounding bands represent 95% confidence intervals. Black jittered dots represent raw data points, and beans represent smoothed density curves.



Probability

We found a significant main effect of Certainty on judgements of probability $F(1, 196) = 6.509$, $p = .011$, $f = 0.18$. Participants who read a letter of diagnosis that was uncertain, rated the probability that the patient has the condition as significantly less likely (mean = 63.1) than when a certain diagnosis was being communicated (mean = 70.0). There was no significant main effect of Face threat ($F(1, 196) = 0.353$, $p = .553$, $f = 0.04$), and no interaction between Face threat and Certainty ($F(1, 196) = 1.019$, $p = .314$, $f = 0.07$).

Certainty

We found a significant main effect of our Certainty manipulation on judgements of perceived certainty $F(1, 196) = 30.483$, $p < .001$, $f = 0.39$. Participants who read a letter of diagnosis that was uncertain, perceived the specialist to be significantly less certain in their diagnosis (mean = 56.3) than when a certain diagnosis was being communicated (mean = 75.0). There was no main effect of Face threat ($F(1, 196) = 0.468$, $p = .495$, $f = 0.04$), and no interaction between Face threat and Certainty ($F(1, 196) = 1.688$, $p = .195$, $f = 0.09$).

Directness

There was a significant main effect of Certainty on judgements of directness $F(1, 196) = 28.735$, $p < .001$, $f = 0.38$. Participants who read a letter of diagnosis that was uncertain, perceived the specialist to be significantly less direct in their diagnosis (mean = 59.8) than when a certain diagnosis was being communicated (mean = 79.0). There was no significant main effect of Face threat ($F(1, 196) = 1.414$, $p = .236$, $f = 0.09$), and no significant interaction between Face threat and Certainty ($F(1, 196) = 0.350$, $p = .555$, $f = 0.04$).

Politeness

There was no significant main effect of Face threat ($F(1, 196) = 1.051$, $p = .307$, $f = 0.08$), and no significant main effect of Certainty ($F(1, 196) = 0.515$, $p = .474$, $f = 0.05$) on judgements of politeness. Additionally, there was no significant interaction between Face threat and Certainty ($F(1, 196) = 0.531$, $p = .467$, $f = 0.05$).

Exploratory Analysis

To determine whether participants in Part 2 were sensitive to linguistic variation in the Part 1 letters (irrespective of experimental condition), we conducted a series of bivariate correlations (see Appendix 4 osf.io/tkxb9). These revealed that the number of hedges was negatively associated with ratings of probability, certainty and directness and positively associated with politeness. Word count was also positively associated with politeness.

4.1.9. Part 2 Discussion

Contrary to our predictions, letters that were written under high face threat were not perceived any differently to letters written under low face threat. Our part 1 face threat manipulation, whereby we instructed participants to address their letter of diagnosis directly to the patient or to the patient's family doctor, did not alter judgements of (1) probability that the patient had the condition (which replicates findings reported by Holtgraves and Perdew, 2016), (2) how certain the specialist was in their diagnosis, (3) how direct the letters were, or (4) how polite the letters were. In short, our part 1 face threat manipulation did not lead to differences in language production that affected the interpretation of participants in part 2. Despite the absence of between group differences, exploratory analysis did indicate that the production of hedges (which can be used to manage negative face) was associated with perceived politeness; letters containing more hedges were perceived as more polite.

In contrast to face threat manipulation, our manipulation of certainty (certain vs. uncertain) in part 1 did alter participants' interpretation on three measures during part 2. Two of these findings reflected the simple fact that uncertain writers produce more uncertain language. First, when letters were written in the uncertain condition, the probability of the patient having the condition was perceived to be lower than letters written in the certain condition. This demonstrates that participants in part 2 could detect this uncertainty via the language of the speaker and lowered their probability estimate as a result. Second, when letters were written in the uncertain condition, the specialist was interpreted as being less certain in their diagnosis. This demonstrates that hearers could detect uncertainty via the language produced by the speaker.

Perhaps most interestingly however, uncertain letters were rated as less direct than certain letters. In other words, the presence of uncertainty made letters appear more indirect than when letters were certain. This demonstrates that the presence of uncertainty when communicating threatening information causes the speaker to produce language is more indirect, and importantly, that this indirectness can be detected by the reader.

4.1.10. General Discussion

The two-part experiment reported in this paper builds upon the work of Holtgraves and Perdew (2016). They found that messages communicated when face management was a concern (high face threat) were perceived as more indirect. Here, we support and extend these findings by evidencing that writers primarily manage face threat using dispreferred markers (e.g., ‘Unfortunately’, ‘I’m sorry to tell you’). These markers alert the reader to the presence or occurrence of threatening news and are an explicit expression of emotional acknowledgement from the writer.

Contrary to our predictions, the face threat manipulation did not affect production of linguistic hedges. Whilst the production of uncertainty terms was found to be infrequent in the data reported by Holtgraves & Perdew (2016), they used relatively informal interpersonal scenarios in their research. This may have encouraged participants to use subtle indirectness as a face-saving strategy, rather than explicitly using uncertainty terms. We hypothesised that increasing the severity of the information to be communicated (i.e., breaking bad news), making the communication more formal (i.e., writing a letter) and increasing relationship distance (i.e., professional rather than personal), would encourage participants to produce hedging terms (e.g., 'possibly', 'maybe') to manage face. Even under these conditions the number of hedges produced did not differ as a function of face threat. Descriptive statistics indicate that those under high face threat did produce slightly more hedges, but the effect size was small. This small, marginally non-significant effect may reflect that our letter writing paradigm allowed participants to edit their letters with no immediate social pressure. Future research that increases the social pressure further (e.g., communicating face-to-face role play) may amplify the face threat and result in larger effects.

This study is the second open-ended language production task to show that writers do not primarily manage face threat with explicit hedging terms. We must therefore consider the possibility that terms like 'possibly' and other hedges are primarily used to communicate genuine uncertainty, rather than as a politeness strategy. Part 1 participants produced many more hedging terms when instructed to deliver an uncertain diagnosis compared to certain. Of course, it makes sense that participants produced uncertainty terms to communicate that they are uncertain. It seems that a writer is more willing to hedge when they believe the diagnosis could be in question (e.g., saying possible when they believe it is probable) compared to when they are convinced of the diagnosis (e.g., saying possible when they believe it is certain).

However, given that our experiment and the experiments reported by Holtgraves and Perdew (2016) did not observe a change in spontaneous production of hedges as a function of face threat, it seems that politeness tactics to manage face in high-stakes situations may manifest themselves in other, more subtle ways.

One of the primary ways writers can use subtle forms of politeness to manage face, is to be indirect. Upon discovering that their participants infrequently produced hedges (e.g., 'possibly', 'maybe') when under face threat, Holtgraves and Perdew (2016) presented evidence that writers may instead manage face by using increased levels of indirectness. For example, they coded phrases like 'you're partying too much' as direct, whereas a phrase like 'Well I'd be failing too if I partied that often' was coded as indirect. They found that face threat resulted in the production of more indirect language, suggesting that rather than achieving politeness via the use of hedges, writers can tactfully 'tiptoe' around sensitive information by being subtly indirect. Critically, the data reported in our experiment also support this claim. First, our participants in part 1 'stretched' the language of their letters with dispreferred markers. By explicitly acknowledging the severity of the situation (i.e., 'this must be tough for you'), and expressing remorse (i.e., 'I'm sorry to inform you'), participants avoided a direct and efficient diagnosis statement which would threaten the reader's positive face. Moreover, these markers encourage the reader to infer the onset of bad news ahead of time. A statement like 'There's no easy way to say this…' for example, indirectly hints towards unfavourable news and causes readers to pre-empt the prognosis. As such, indirectness is used to achieve politeness under face threat through the use of dispreferred markers. Second, part 1 participants chose to use significantly more words to complete their letter while under face threat (which likely reflects the increased use of dispreferred markers). One plausible explanation for this finding is that the expected content of the letter is different

when addressed to a patient rather than a doctor, based on presumed responsibilities and expertise. For example, stating the purpose of a particular health assessment (e.g., a brain scan) may be useful when writing to a patient, but wholly unnecessary when writing to a fellow doctor (and perhaps even impolite or patronising). Despite this, the increase in word count could also be explained by the aforementioned 'polite indirectness'. It is often the case that the politest phrasing deviates from perfectly efficient communication (e.g., saying 'if you could pass the salt, that would be great', rather than 'pass the salt'; Pinker et al., 2008). To choose speech that communicates in this way is to speak indirectly. Based on our word count data, it appears that writers under face threat demonstrate a reluctance to deliver undesirable information. This reluctance has previously been referred to as the 'MUM effect' (Rosen & Tesser, 1970). It appears that writers in our experiment chose to indirectly attend to the reader's face by 'padding out' their letters with additional (unnecessary) words.

Manual inspection of the letters suggests that 'stretching' occurred in other ways, such as giving a full description of the disease, or citing support helplines. It's possible that these are techniques characteristic of a 'forecasting' or 'stalling' approach to breaking bad news (Shaw et al., 2012). Observed in doctors, Shaw and colleagues detail that a 'forecasting' approach represents a staged delivery of bad news (within the first 2 minutes), whilst a stalling approach represents an outright postponement of the delivery (beyond 2 minutes). The additional (or unnecessary) content used to 'pad out' letters in part 1 may reflect these delivery styles, potentially opening an avenue of further investigation. Researchers may, for example, wish to conduct exploratory analysis using our open access data.

We support the interpretation of Holtgraves and Perdew (2016), that face threat can be subtly managed via indirect language. This is a particularly notable finding. We know that tactful indirectness can be beneficial in everyday social situations that threaten face, such as

attempting to get out of a speeding ticket or expressing a negative opinion to a friend (Pinker et al., 2008; Sirota & Juanchich, 2015). However, this research extends our knowledge, demonstrating that language producers employ subtle politeness tactics even in particularly high-stakes situations (i.e., breaking bad news to patients). Even under these extreme circumstances, it seems that face threat is managed through subtle indirectness.

One of the most notable effects in part 1 was the time taken to compose the letters. The effect of face threat (high vs low) on writing time only occurred when the diagnosis was uncertain. There was no such effect when the diagnosis was certain. This gives a novel insight into how managing both uncertainty and threat plays out in the mind of the writer in real-time. We present a potential explanation for this finding. Having to balance both uncertainty and sensitivity at the same time is particularly difficult to achieve. Bonnefon et al. (2011) evidence that disentangling politeness is taxing on mental resources, meaning that greater cognitive effort is needed to 'read between the lines'. Presumably, placing speakers in situations that require them to use politeness, creates somewhat of a 'barrier' to efficient communication. When faced with the balancing act of accurately communicating their level of commitment to the statement, whilst attending to the face of all parties, writers are forced to use greater cognitive resources to overcome these linguistic hurdles put forward by the circumstances. It is entirely plausible therefore, that the cumulative challenges posed by uncertainty and threat together, accounts for an increase in time spent writing the letter.

In part 2 of this experiment, we examined the comprehension of the letters produced in part 1. This was based on a method by Holtgraves and Perdew (2016) and gives additional insight into the communicative dyad. Beginning with the effect of our certainty manipulation, we found that letters written in our 'certain' condition in part 1 were perceived to be more certain by participants in part 2. Likewise, letters written in the 'certain' condition were

perceived as more direct and to be communicating a higher probability that the patient truly has the diagnosed condition. This suggests that our experimental manipulation of certainty in part 1 produced changes in language production that were detected by participants in part 2. In general, our manipulation of the writers' certainty is particularly valuable here in achieving a more complete understanding of uncertainty communication, primarily because it allowed us to observe how uncertainty is handled in particularly threatening circumstances.

Contrary to our predictions, letters written under high and low face threat were not judged differently on any of our interpretive measures (Probability, Certainty, Directness, Politeness). Despite letters written under high face threat in part 1 containing more words and more dispreferred markers than letters written under low face threat, these letters were not actually perceived as any less direct by participants in part 2. Likewise, the change in language use caused by our part 1 face threat manipulation did not affect how participants in part 2 perceived the certainty of the diagnosis or the probability that the patient truly has the diagnosed condition. From one perspective, these null findings are reassuring, as they suggest that facework strategies do not cloud the core meaning of the message being communicated. Interestingly, this aligns with a wider body of research evidence suggesting that despite the multiple possible motivations for speech, humans are particularly good at mitigating the use of face-management strategies and disentangling ambiguity (Aronsson & Sätterlund-Larsson, 1987; Bromme et al., 2012; Person et al., 1995). For instance, hearers seem to be good at knowing that the hedged polite utterance 'I don't think you're suffering from measles' in practice means 'Your concerns are unwarranted, this isn't measles' (Brummernhenrich & Jucks, 2019). In other words, the hearer catches the intended meaning of the message despite the ambiguity created by politeness. From another perspective though, the null findings could

suggest that the effort made by language producers to manage face threat may be in vain, as it is not actually picked up by recipients.

Participants in part 2 were also asked to judge the politeness of the letter and we found that politeness ratings did not differ as a function of certainty or face threat. We believe there may be an elementary explanation for our null finding. Participants in part 2 were not asked to place themselves in the position of the recipient (either the patient or family doctor), rather, they read the exchange from the third person (i.e., from the outside looking in). It is therefore possible that context-dependent expectations of politeness confounded the judged level of politeness between the two scenarios. For example, a relatively terse letter written by a medical specialist to a family doctor may be considered polite in that professional context, even though that same letter might not be considered polite if it were addressed to the patient. This is because changing the addressee of the letter (i.e., from doctor to patient) changes the degree of politeness required to manage the threat of that situation, and by extension, changes the type of politeness that one would expect in that situation. This could mean that our participants in part 2 judged relative (contextual) politeness, rather than absolute (objective) politeness. Because of this, when they were asked 'how polite is this letter?', participants are likely to have judged the letters in their respective context, because they knew the intended recipient of the letter. This means that our letters may not be equally polite in absolute terms, but when judged within their own individual circumstances they are rated as equally polite.

Another consideration with regards to ratings of politeness in part 2, is that in addition to facework, participants may have also been influenced by indicators of deference (e.g., please, thank you). These linguistic markers are not used explicitly to manage face, but when used in discourse, they are commonly attributed to a person that is being polite. Indeed, the

discrepancy between the academic vs. everyday understanding of politeness may have had a non-trivial influence in this sample; and should be considered as a potential limitation.

We must also consider other limiting factors and their influence on the effects (or lack of effects) outlined above. First, some of the effect sizes observed in part 1 and part 2 were small whereas our study was designed to detect only 'medium' sized effects. Second, we did not perform manipulation checks in part 2, so cannot verify whether face threat was effectively operationalised within our scenarios. Differences in wording between the conditions were subtle (i.e., non-explicit), so a potential instance of the letter addressee being overlooked (e.g., a participant assuming that the letter is addressed to the patient, rather than the doctor) would not have been detected in the analysis.

The communication of uncertainty is a complex linguistic process, particularly under circumstances that threaten the positive social identity of the parties involved. We tested the production and comprehension of uncertain language within a health context, manipulating how socially delicate (or face threatening) the situation is. Our results indicate that writers manage threat under difficult social circumstances by using dispreferred markers (e.g., 'I'm sorry to inform you', 'I understand', 'this must be a difficult time for you') to acknowledge the damaging effects of the threatening information, and to address the sensitivity of the situation. These face-saving strategies used by language producers did not appear to cloud comprehension of the core messages being communicated.

Acknowledgements: We thank Molly Wheeler and Isabel Dunkerley who acted as independent blind coders in part 1 of this study.

Author Contribution Statement

Study Conception: Harry Clelland, Matthew Haigh

Study Design: Harry Clelland, Matthew Haigh

Development of Materials: Harry Clelland

Pre-Registration: Harry Clelland

Data Collection: Harry Clelland

Data Analysis: Harry Clelland

First Draft: Harry Clelland

Review and Editing: Harry Clelland, Matthew Haigh

4.2. Chapter Summary

The experiments reported in this chapter set out to answer our first research question: '*How do pragmatic factors influence the production and interpretation of uncertainty terms?*'. We addressed this question with a large two-part experiment, testing whether uncertainty and face-threat would have an effect on the production and comprehension of non-referential terms such as 'perhaps', 'maybe', 'could'. Below we consider how the findings reported in this chapter directly answer our research question (RQ1).

Predictably, uncertainty had an effect on the number of hedging uncertainty terms used, which simply reflects the fact that an uncertain speaker will hedge their commitment to a proposition by expressing explicit uncertainty (e.g., you have '*possibly*' been diagnosed with Alzheimer's disease). However, we found no effect of face-threat on hedging uncertainty terms, suggesting that these terms are used primarily to communicate genuine uncertainty, rather than to achieve face-work. Face-threat had a significant impact on the number of dispreferred markers used (e.g., 'unfortunately'), whereby this type of language was used more often under high face-threat. Participants used more overall words under high face-threat, and they took significantly longer to produce language when under both high face-threat and uncertainty. Taken together, our findings advance knowledge by evidencing that, in line with the findings of Holtgraves and Perdew (2016), face-threat does not influence the production of explicit uncertainty terms (e.g., 'possibly'). Instead, face-threat significantly impacts indirectness, whereby more threatening circumstances produce more indirect language. In part 2 we found that the only detectable changes in language from part 1 came from uncertainty, and not from high face-threat. This suggests that increasing the threat of a situation does not necessarily cause differences in language production that are detectable by

hearers. However, introducing uncertainty causes hearers to perceive a lower probability and a less certain speaker.

In the following chapter, we test the pragmatic factors that influence two types of quantifying language. First, we look at proportional quantifiers (e.g., 'some'). These terms are similar to uncertainty terms in that they do not refer to a solid proportion (Holtgraves, 2014). However, terms like 'a few' or 'many' are far more relevant within the context of health communication because they are preferred by family doctors when communicating risk (Juanchich & Sirota, 2020a). Second, we look at '1-in-X' ratios (e.g., '1 in 7 chance of suffering side effects'). At face value, these ratios seem less ambiguous than proportional quantifiers because they specify an exact likelihood (e.g., 14.3%). But, they are subject to a cognitive bias that reliably causes an overestimation of risk (Pighin et al., 2011; Pighin et al., 2015; Sirota & Juanchich, 2019). Chapter five explores how changes to context influence the use and perception of both verbal and numeric quantifiers (RQ2).

Chapter 5. How do Pragmatic Factors Influence the Production and Comprehension of Quantifiers?

5.1. "Some patients suffer…" – Proportional Quantifiers are used Tactfully (Exp. 3)

5.1.1. Abstract

When communicating uncertain negative information, a speaker can politely soften the threat of this information by altering explicitly communicated probability (e.g., saying a negative outcome is 'possible' when they believe it is actually 'probable'). The aim of this study was to establish whether speakers also manage face threat by using proportional quantifiers such as 'some' or 'many'. In this experiment, participants imagined themselves as a health specialist and communicated a negative uncertain medical opinion (an anticipated diagnosis). This was communicated either directly to the patient (high face-threat) or to the patient's family doctor via letter (low face-threat). Participants were asked to choose a proportional quantifier to express their commitment to the diagnosis (either 'a few patients', 'some patients', 'many patients' or 'most patients'). Under high face-threat, participants were more likely to choose a lower proportion quantifier (e.g., 'some' rather than 'many') and more likely to report that their communicative intention (motivation) was face-management (rather than to be informative). We demonstrate that in the same way that uncertainty terms (e.g., maybe) and probabilistic terms (e.g., likely) are used to soften threatening news, so too are proportional quantifiers. When used to communicate about health, the non-referential nature of proportional quantifiers mean that they have the potential to cause misunderstandings between doctor and patient.

5.1.2. Introduction

Goffman's theory of 'Face' postulates that each individual possesses their own public self-image, which is sacred to the individual, and as such, requires active protection (Goffman, 1967). During social interaction, a person voluntarily places their social self-image at risk. For example, imagine that you were recently in the passenger seat of a car, and were involved in a minor accident. You think that the accident was caused by your friend, who was texting and driving at the time of the crash. Later, when you are alone with your friend they ask you about what you think caused the crash. In this example from Holtgraves and Perdew (2016), expressing an honest opinion to your friend about the cause of the crash poses a risk to face. That is, telling them that they are the cause threatens their positive social self-esteem. In these situations, a person is likely to engage in 'face-work' (Goffman, 1971), which essentially refers to any action sought to dampen the threatening nature of a speech act.

Considered a direct extension of Goffman's work, Brown and Levinson's (1978; 1987) Politeness Theory posits that face-work is achieved through politeness-based strategies. Put simply, a person can protect their own face (and the face of their communicative partner) by being polite. One particularly important mechanism that manifests when a person intends to be polite, is that they will deliver information that differs from the information that they actually wish to communicate. This is particularly the case when the negative information being communicated is uncertain. For example, consider again your friend who has asked for your opinion about the cause of the recent car crash. The genuine cause of the crash is inevitably uncertain, because whilst you were witness to your friend's texting behaviour, it may also be the case that the driver of the other car was at fault. So, when communicating your negative opinion to your friend you might say that it's 'possible' (rather than 'likely') that they caused

the crash, effectively altering the information you had in mind in the interest of preserving face.

A number of research papers have used this proposition of politeness theory (described above) to hypothesise about how face-threatening messages would be interpreted by a hearer (i.e., the person on the receiving end of the message). Bonnefon and Villejoubert (2006), for example, discovered that the severity effect (the bias for severe events are judged as more likely to occur) can be explained by a hearer's expectations of politeness. In their data, participants considered '*possible deafness*' to be more likely than '*possible insomnia*', and part of the reason for this was because when the outcome was severe (deafness), participants considered the term 'possible' to be a politeness strategy. That is, hearer's inflated their subjective ratings of probability, under the assumption that the person delivering the news would have attempted to soften their genuine negative outlook. A similar pattern of results is observed for scalar inferences. The tendency for the word '*some*' to be interpreted to mean '*some, but not all*' changes depending on whether or not the speaker is considered to be communicating tactfully (Feeney & Villejoubert, 2013). Lastly, this effect is again observed for interpretations of verbal or numerical quantifiers. Across five experiments, Juanchich et al. (2012) evidenced that when speakers were assumed to have face-management motivations, hearers would adjust the numerical meaning of probabalistic terms like '*small probability*' or '*quite likely*' upwards, to compensate for the fact that the speaker is likely downplaying their genuine beliefs.

It is clear that research on uncertain communication and politeness is plentiful from the perspective of the hearer (language comprehension). Less frequent though, are studies that observe communicative intentions (motivations) and communicated probability from the perspective of the speaker (language production). To address this issue, Sirota and Juanchich

(2015) tested whether a person's communicative intentions could influence how they choose to communicate uncertainty. Participants were given a hypothetical scenario, in which they were tasked with communicating a negative outcome at 50% probability (uncertain). In one scenario for example, the speaker expresses to their friend the possibility that their financial investments will lose value. Participants selected a verbal probability to communicate the bad news, ranging from a relatively low likelihood ('very small probability') to a relatively high likelihood ('quite probable'), and a centre-point on the scale that represented 50% likelihood ('evenly probable'). After this, they reported their communicative intentions as either informative or face-management. In other words, participants either intended to communicate the correct probability, or they intended to communicate in a way that manages the face of either themselves or the hearer. The analysis revealed that participants altered their communicated probability more often and more substantially when adopting face management techniques. Specifically, those who were motivated by caution and tact (face-management) communicated a significantly lower average probability than those who were motivated to be informative.

The findings of Sirota and Juanchich (2015) are enlightening because it was shown via direct empirical evidence that a speaker will alter the communicated probability of uncertain bad news when motivated by face-management intentions. Before this point, there existed evidence only from the perspective of the hearer, that a speaker is thought to 'soften the blow' of bad news by hedging and downplaying likelihood (Bonnefon et al., 2011a; Bonnefon & Villejoubert, 2006; Feeney & Villejoubert, 2013; Juanchich et al., 2012). Put simply, a speaker who is looking to manage face will overstate or understate the likelihood of an event, depending on whichever achieves face-work.

A key question that remains following this discovery, is whether the tendency to alter communicated probability is emphasised under high face-threat compared to (relatively) lower face-threat. Specifically, is it the case that when faced with a particularly threatening situation, a person will (1) communicate an altered probability and (2) report face-management motivations more often, compared to the same situation with experimentally lowered face-threat. Based on empirical findings it is reasonable to argue that this will be the case. Across two experiments, Holtgraves and Perdew (2016) examined (via hypothetical scenarios) how participants would choose to communicate the likelihood of negative outcomes. Their first experiment evidenced that, all else being equal, speakers would communicate the probability of a more severe outcome to be less likely to occur than a less severe outcome. For example, the communicated likelihood of a broken down car needing a new transmission (severe) was lower than the same car needing a new battery (non-severe). In their second experiment, the authors manipulated face-threat by keeping the scenario identical but changing the referent. For example, a negative opinion would be communicated whilst the subject of that opinion was either present (high face-threat) or not present (low face-threat). Participants used indirect language as a politeness tactic under high face-threat, and this language was subsequently interpreted as more uncertain (lower probability). Crucially, the findings of both experiments demonstrate that communicated probability is altered by speakers in severe situations, and that a lower communicated probability is achieved through politeness strategies under high (rather than low) face-threat. In the experiment reported below, we build on the work of Sirota and Juanchich (2015), as well as the subsequent work of Holtgraves and Perdew (2016), by applying the study of uncertain language production to the communication of health risk. In particular, the production of proportional quantifiers such as ‘some’ or ‘many’.

5.1.3. The Present Research

The aim of this experiment was to extend the findings of Sirota and Juanchich (2015). They discovered (in line with a key postulate of politeness theory) that a speaker will alter the communicated likelihood of a negative event when motivated by face-management strategies. Here, we aim to establish whether an explicit manipulation of face-threat would alter motivations at a more fine-grained level. In other words, does a speaker tap into face-management tactics for generally threatening outcomes, or does the extent to which this happens depend on precisely *how* threatening the situation is? One of the ways in which we aim to answer this question is by heightening the 'stakes' of the hypothetical scenario, asking participants (within a hypothetical medical context) to 'break bad news' to a patient (Baile et al., 2000; Burgers et al., 2012; Shaw et al., 2011). Delivering a diagnosis under the more formal (less interpersonal) circumstances surrounding doctor-patient communication, is more threatening than communicating a negative opinion to a friend, for example, meaning our resulting manipulation of face-threat can be considered stronger than that of previous works.

In the current study we also make an important adaptation to the type of language used to communicate likelihood. Unlike Sirota and Juanchich (2015), who provided participants with a scale of probabilistic terms (e.g., 'quite likely'), here we will test if findings extend to proportional quantifiers (e.g., 'a few', 'some'). This is important to test within the communication of uncertain health information because the use of quantifiers is one of the most popular ways for family physicians to communicate adverse risk to their patients (e.g., '*a few patients suffer…*'; Juanchich & Sirota, 2020a). As such, when communicating negative health information, participants chose from a set of four quantifying terms ranging from low proportion (a few) to high proportion (most), based on mental representation data from Pezzelle et al. (2018a).

In one of several hypothetical scenarios created for this study, participants were asked to imagine themselves as a healthcare specialist (a specialised doctor, rather than a family doctor or general practitioner). They receive information about a patient's recent symptoms, as well as the results of a test recently conducted to figure out the potential health issue. Participants are told that based on their expertise, they are between 45% and 55% sure of a potential diagnosis (i.e., uncertain), and are then tasked with communicating this to either the patient in-person (high face-threat), or to the patients' GP (family doctor) via letter (low face-threat). Researchers have demonstrated that manipulating face-threat in this way (i.e., changing the referent) is an effective methodology (Holtgraves, 2014; Holtgraves & Perdew, 2016).

As a function of their given scenario, participants selected a proportional quantifier that represents their confidence in the diagnosis (either '*a few patients*', '*some patients*', '*many patients*' or '*most patients*'). On this measure, we predicted that the distribution of quantifier choice would differ as a function of face-threat. Finally, participants told us their primary motivation when choosing a quantifier. That is, when choosing how to communicate the potential diagnosis, participants were either motivated to be informative (accuracy), or they were motivated by either speaker or hearer face-management (caution, tact). Here we predicted that the distribution of participants' motivations would differ under high face-threat compared to low.

5.1.4. Method

This experiment was pre-registered prior to data collection (see osf.io/kwpnm). Raw data, analysis script and materials are freely available on the Open Science Framework (see osf.io/njw54/). Our aim in this study was to test if heightened face-threat would impact (1) choice of a proportional quantifier (e.g., some, many) used to communicate uncertainty, and (2) the communicative intentions that motivated participants' choice of quantifier. Respondents imagined themselves in the position of a medical professional (i.e., a specialist in a particular condition), and are tasked with communicating a suspected diagnosis, of which they are uncertain (45%-55% certain). When face-threat was low, this opinion was communicated via letter to the patient's GP (family doctor), whereas when face-threat was high, the opinion was communicated directly (in-person) to the (hypothetical) patient themselves. Based on these circumstances, participants indicated a proportional quantifier that they would use to communicate their opinion, and reported their primary intention when making this decision.

Design

This one-factor independent groups design made a single manipulation. Face-Threat differed across two levels through a change in both the referent and the communicative format (high = communicating to the patient in-person, low = communicating to the patient's GP via letter). We measured participants' choice of proportional Quantifier (DV1) used to qualify an uncertain medical opinion (either 'a few patients', 'some patients', 'many patients' or 'most patients'). We also measured participants' Motivation (DV2) for choosing their specific quantifier (either Informativeness, Hearer Face-Management or Speaker Face-Management).

Participants

We conducted a prospective power analysis using the 'pwr' package in R (Champely, 2020). Based on the three post hoc tests that we pre-registered, we required a minimum 156 participants, per test, to detect a medium effect size 80% of the time (using α = .0167). As an approximation, each test utilised ⅓ of the entire sample, therefore, we required a minimum of 468 participants (156 x 3).

An initial sample of 497 consenting volunteers were recruited via opportunity sampling methods (social media, university mailing lists). We removed the responses of those who did not complete the study (n = 22) and those who openly declared a non-serious response (n = 1). Our final sample was made up of 474 English-speaking adults (118 men, 345 women, 9 non-binary, 2 self-identify) aged between 18 and 66 years (mean age = 33.22, SD = 11.85). The number of participants randomly allocated to the high face-threat condition was 238, and the number of participants allocated to the low face-threat condition was 236.

Materials

We created a total of six experimental vignettes for this study. Each vignette places participants into a hypothetical scenario, whereby they imagine themselves as a health professional (a specialist in a specific condition). They are given information about a fictional patient from their family doctor (known as 'general practitioner' or GP in the UK), along with a number of recently reported symptoms. Participants are provided with the patient's recent test results, and told that based on their expertise, they are between 45% and 55% confident that the patient is to be diagnosed with a disease (uncertain). This uncertain diagnosis was either communicated to the patient's family doctor (GP) via letter, or to the patient directly in-

person. An example scenario can be seen in Figure 7, and all six scenarios in full can be found on the OSF (see osf.io/p5nj4).

Participants were allocated to one of 12 vignette versions (see Table 3). To aid the believability of scenarios, the chosen diseases have no definitive test, and therefore require some level of professional opinion. There were two experimental conditions embedded within each scenario, resulting in a total of 12 vignettes. Participants were randomly allocated to one vignette. The title of the fictional patient (Mr or Mrs) was counterbalanced across scenarios.

Figure 7

Example scenario. The case of Mr Wilson, to be (potentially) diagnosed with Type II Diabetes.

Imagine you are a specialist doctor at a medical practice. You receive some information about a patient, Mr Wilson, from their family doctor (GP).

Mr Wilson has been to his doctor a number of times in recent weeks reporting dizziness and trouble with sleep. The GP took a sample and has sent it to you for analysis. His blood sugar levels appear to be high at 10.2 mmol/L.

You know based on your expertise that a patient with these results has between a 45% and 55% chance of being diagnosed with Type II diabetes. For this reason, you are going to [write a letter to Mr Wilson's GP/have a consultation with Mr Wilson] and recommend that he has more extensive testing.

TASK - Complete the passage below

[You are writing a letter to Mr Wilson's GP about his results/Mr Wilson comes into your practice to get his test results]. You [write/say]…

"A few patients with these test results have Type II Diabetes, so I am going to recommend further tests"
"Some patients with these test results have Type II Diabetes, so I am going to recommend further tests"
"Many patients with these test results have Type II Diabetes, so I am going to recommend further tests"
"Most patients with these test results have Type II Diabetes, so I am going to recommend further tests"

Table 3

Scenario, fictional patient name and potential disease as a function of low and high Face-Threat conditions. 12 total vignettes (6 Diseases x 2 Conditions).

High Face-Threat			**Low Face-Threat**		
Scen.	**Patient**	**Disease**	**Scen.**	**Patient**	**Disease**
1	Mr Wilson	Type II Diabetes	1	Mr Wilson	Type II Diabetes
2	Mr Wilson	Coronary Heart-Disease	2	Mr Wilson	Coronary Heart-Disease
3	Mr Wilson	Dementia	3	Mr Wilson	Dementia
4	Mrs Wilson	Parkinson's Disease	4	Mrs Wilson	Parkinson's Disease
5	Mrs Wilson	Thyroid Disease	5	Mrs Wilson	Thyroid Disease
6	Mrs Wilson	Coeliac Disease	6	Mrs Wilson	Coeliac Disease

Face-Threat Manipulation

To operationalise the level of face-threat within our scenarios, we created both a low face-threat and a high face-threat version. In the low face-threat version, participants are tasked with communicating their negative health opinion via a letter to the patient's family doctor (GP). In the high face-threat version, participants are tasked with communicating their negative health opinion to the patient themselves in a face-to-face consultation. Heightening face-threat between the two conditions is achieved here in two ways. First, delivering the news directly to the patient (rather than indirectly via the GP) is more threatening because it is a direct challenge to the patient's positive self-image. More than this, breaking bad news to a patient directly is more face-threatening for the participant themselves, described as "*dropping a bomb*" by doctors, alongisde feelings of guilt and responsibility (Miyaji, 1993). Second, communicating face-to-face (rather than via letter), even in a hypothetical capacity is more face-threatening because the speech act is more imposing on the hearer. Politeness Theory

(Brown & Levinson, 1987) dictates that degree of imposition (Rx) increases the 'weightiness' of a speech act, and an increase in imposition is associated with greater use of politeness tactics (Lambert, 1996; Okamoto & Robinson, 1997). Considering both of these factors, we consider our manipulation here to be an effective operationalisation of face-threat within our vignettes.

Measures

Proportional Quantifier Choice

After reading the scenario, participants are tasked with choosing a speech act to communicate an uncertain negative medical opinion, from a list of four options. Each option was identical, except for the proportional quantifier used in the passage. Participants select one of four statements reading "*[A few, Some, Many, Most] patients with these test results have [disease], so I am going to recommend further tests*". The order-of-presentation was randomised for each participant.

Motivation

Communicative intentions (motivation for choosing the quantifier) was measured by asking participants "*What was your primary intention when choosing a statement?*". Participants selected one of three possible options. The first was "*I wanted to communicate the probability*" which indicated an intention to be informative (informativeness). The second was "*I tried to give my opinion in the most tactful way*" which indicated an intention to engage in Hearer Face-Management (to protect the face of the patient). Finally, the third was "*I tried to be cautious, in case I could be wrong*" which indicated an intention to engage in Speaker Face-Management (to protect the participant's own face). The order-of-presentation was randomised for each participant.

This measure mirrors that of Sirota and Juanchich (2015), however, in their experiment participants were given a fourth option in which they could enter any other motivations they may have had. It was discovered that this option was used infrequently by participants, and even when used, freely written responses would often reflect that of one of the original, predetermined options (e.g., '*better safe than sorry*' can be considered a cautious approach, or speaker face-management). Based on this, we considered the three options detailed here to be sufficient in effectively capturing motivation.

Measure Check (Quantifiers)

Based on data from Pezzelle et al. (2018a), we reason that the actual proportion represented by each of our quantifiers is ordered. That is, '*a few*' represents a smaller proportion than '*some*' followed in order by '*many*' then '*most*'. Considering our hypotheses, it was important to confirm that this is the case among the current sample. All participants were asked four questions prefaced with "*In your opinion, what <u>percentage of patients</u> is represented by each of the expressions presented below?*". On a sliding scale ranging from '0% of patients' to '100% of patients', participants provided subjective ratings of the proportion represented by '*a few patients*', '*some patients*', '*many patients*' and '*most patients*'. All participants rated each statement, and the order-of-presentation was randomised for each participant.

Procedure

Access to this online study was granted via web link to Qualtrics survey platform (Snow & Mann, 2013). Participants answered basic demographic questions, and following this their response was prefaced with the following text:

"On the following page you will be presented with a short imaginary scenario, in which you will be asked to imagine yourself as a healthcare professional. Please read the scenario and answer the 6 questions that follow, based on your own opinion."

Once instructed, participants were randomly allocated to one of 12 possible vignettes (6 x scenarios, 2 x conditions). As a function of the scenario, participants selected their choice of proportional quantifier, and provided their primary motivation for choosing this quantifier. After this, all participants were directed to our measure check question block, where they indicated the proportion represented by all of our quantifier options (a few, some, many and most). Before submitting their response, participants completed a seriousness check (Aust et al., 2013). Debrief information was provided as part of the end-of-survey message. Median completion time for the entire survey was 2.6 minutes. A .qsf download file for the survey can be accessed via the OSF (see osf.io/xmf3w). This can be used to replicate our method exactly.

5.1.5. Results

The following analysis was pre-registered prior to data collection. The raw data from this study, as well as the script created to analyse the data can be found on the OSF (see osf.io/7jgr5 for data, see osf.io/2kqar for analysis script).

All questions required a response, so there was no missing data and no data points were excluded. Analyses were carried out using R version 4.1.0 (R Core Team, 2021) and RStudio version 1.4.1717 (RStudio Team, 2021). We performed a total of six tests, all of which were Chi-Square (χ^2) test of independence using the 'gmodels' package in R (Warnes et al., 2018). For each test we also calculated Adjusted Standardised (Pearson) Residuals ($\tilde{r}_{ij}$), to determine whether cell counts significantly deviated from their expected counts. Cells that

exceed the absolute value = +/-1.96 indicate a significant deviation from the expected count. Phi (φ) and Cramer's V (V) are reported as measures of effect size.

First, we observed the association between Face-Threat and Quantifier (2x4 test). To provide further clarity on this analysis, we ran an exploratory (not pre-registered) 2x2 test, excluding those who chose 'a few' and 'most'. Then, we observed the association between Face-Threat and Motivation (2x3 test). Finally, we observed the association between Face-Threat and Quantifier as a function of communicative Motivation. This was achieved via three 2x4 tests on subsetted data (e.g., informative responses only).

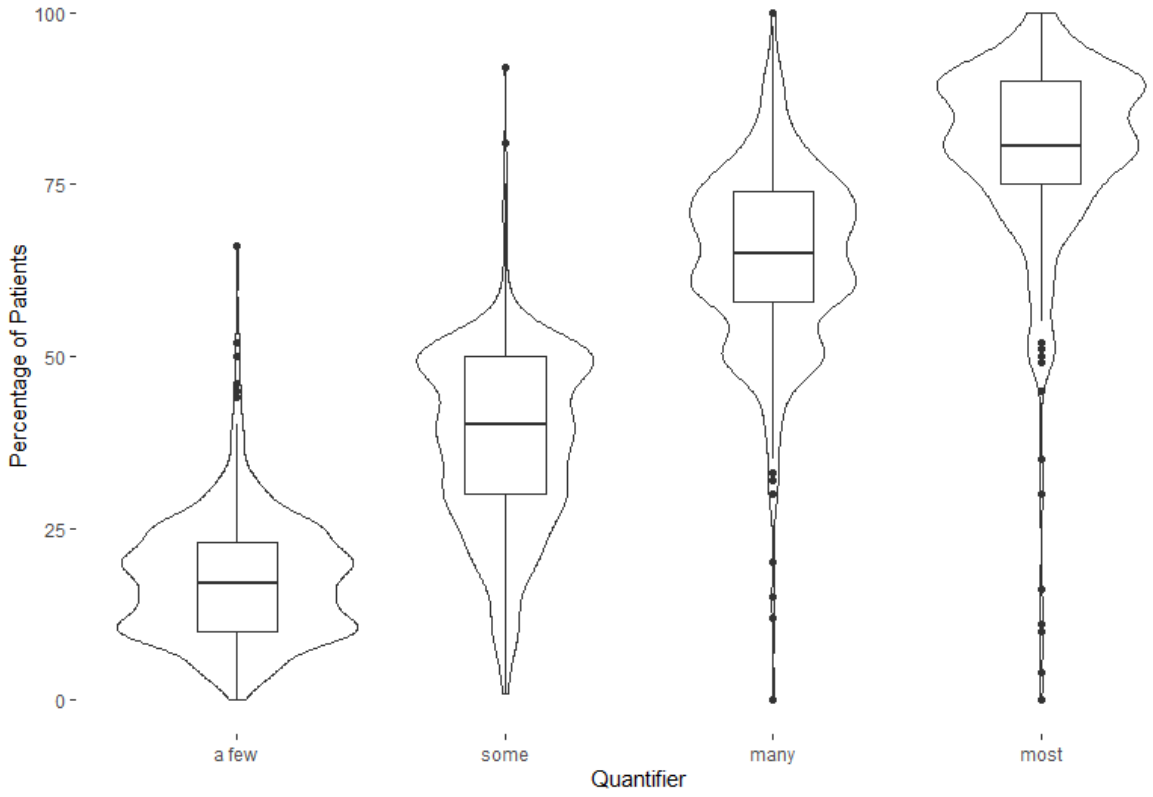
Additional exploratory analysis was conducted to observe whether scenario (i.e., the 6 vignettes) was significantly associated with our DV's. This confirmed that there was no association between Scenario and Quantifier $\chi^2(15, N = 474) = 12.813$, $p = .617$. There was however a significant association between Scenario and Motivation $\chi^2(10, N = 474) = 19.431$, $p = .035$. Participants were motivated to be more informative in scenario 1 (Type II Diabetes, $\tilde{r}_{ij} = 2.566$) and less informative in scenario 4 (Parkinson's Disease, $\tilde{r}_{ij} = -2.054$). Note that this association does not account for the influence of face-threat (i.e., who the participants are being informative *to*).

Measure Check (Quantifiers)

The data presented in Figure 8 shows the proportion of patients represented by each quantifier.

Figure 8

Violin plot showing subjective ratings of the percentage of patients represented by each quantifier (N = 474). Vertical lines represent the median proportion value, and surrounding boxes represent the 25th and 75th percentile (interquartile range). Smooth density curves represent the kernel probability density of the data across a range of values.



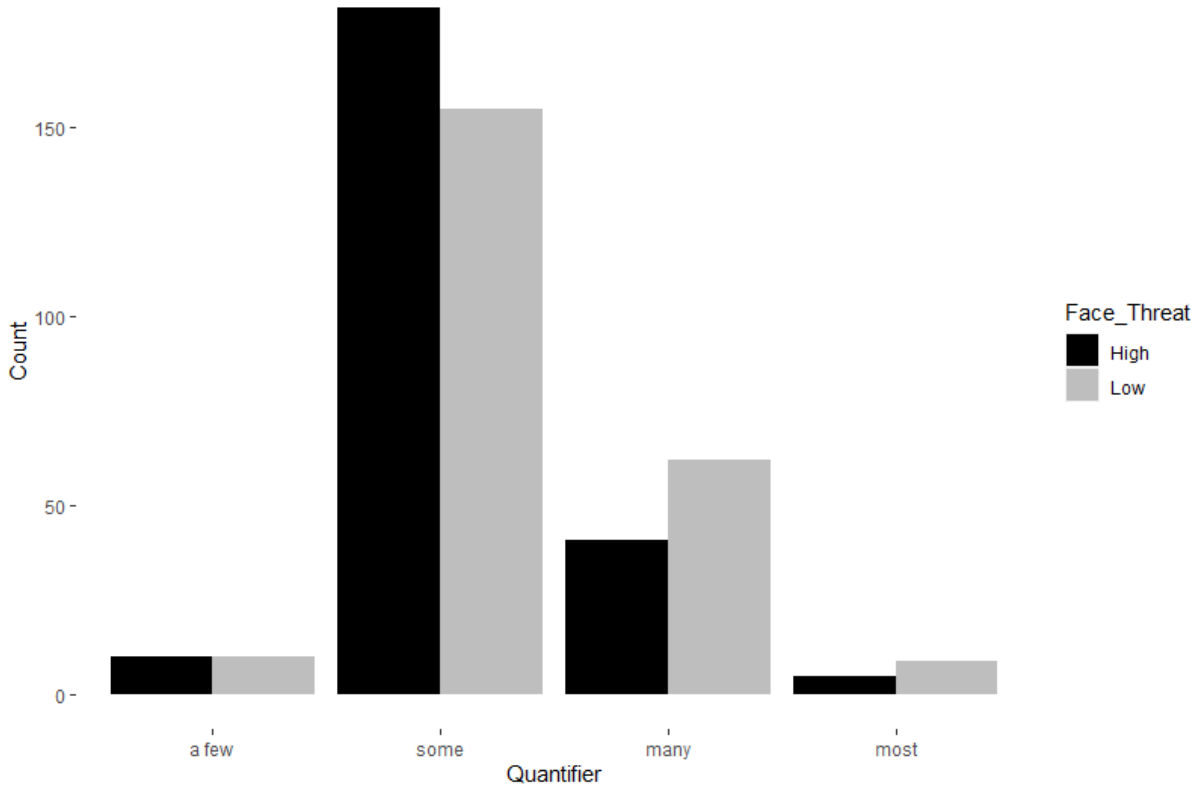
The pattern of results displayed in Figure 8 shows that the proportion magnitude order was preserved. That is, the proportion of patients represented by 'a few' (M = 17.5, SD = 8.7) was smaller than 'some' (M = 37.4, SD = 13.0), followed by 'many' (M = 64.1, SD = 12.9) then 'most' (M = 79.8, SD = 13.7). The difference between all combinations of quantifiers was highly significant ($p < .001$). Unlike Sirota and Juanchich (2015), the scale used in this study does not include a quantifier that represents 50% probability (e.g., evenly probable). However, the data for 'some' and 'many' show that these quantifiers fit nicely around our percentage-of-certainty range (45%-55%).

Association between Face-Threat and Quantifier

After reading the scenario, participants chose one proportional quantifier that they believed was most appropriate to use when communicating an uncertain negative medical opinion. These findings are summarised in Figure 9.

Figure 9

Bar chart showing the association between Face-Threat (high, low) and Quantifier choice (a few, some, many, most). The y-axis represents the raw number of participants who chose each quantifier, as a function of whether they were under high or low face-threat.



Surprisingly, the overall association between Face-Threat and Quantifier choice was marginally non-significant $\chi^2(3, N = 474) = 7.579$, $p = .056$, V = .13. However, upon observing specific cell counts we discovered that between high and low face-threat conditions, the proportion of participants choosing ‘some’ and ‘many’ differed significantly from expected cell counts. When under high face-threat, 76.5% of participants chose to qualify the

negative health opinion with ‘some’. This was a significantly larger proportion than participants under low face-threat (65.7%; $\tilde{r}_{ij}$ = 2.592). The opposite pattern was observed for the use of ‘many’. When under high face-threat, 17.2% of participants chose to qualify the negative health opinion with ‘many’, a significantly lower proportion than participants under low face-threat (26.3%; $\tilde{r}_{ij}$ = 2.387).

Association between Face-Threat and Quantifier (Exploratory, Not Pre-Registered)

We suspected that the reason our original analysis failed to reach statistical significance was due to ‘noise’ created by the inclusion of ‘a few’ and ‘most’ choice options. These options were seldom selected by our participants (4.2% of the entire sample chose ‘a few’, 3% of the entire sample chose ‘most’). We therefore decided to (temporarily) eliminate these responses from our data, and run a 2x2 chi-square test to again observe the association between Face-Threat (high, low) and Quantifier choice (some, many). When analysing the data in this way, we find a significant overall association $\chi^2(1, N = 440) = 6.364$, $p = .012$, $\varphi = .11$, suggesting that as suspected, the inclusion of ‘a few’ and ‘most’ choice options in the original analysis created additional noise in the data, effectively ‘diluting’ the observed significance level.

Association Between Face-Threat and Motivation

After reading the scenario and choosing an appropriate quantifier, participants indicated their primary motivation (communicative intentions) when making their choice. This was one of three options, (1) Informativeness, (2) Hearer Face-Management (tact), or (3) Speaker Face-Management (caution). These findings are summarised in Table 4.

Table 4

Count, expected count and adjusted standardised residual values for the association between Face-Threat (high, low) and Motivation (Informativeness, Hearer Face-Management, Speaker Face-Management).

	High Face-Threat		
	Informative	**Hearer FM**	**Speaker FM**
Count	91	83	64
Expected Count	104.44	77.33	56.24
Adjusted Std. Residual	-2.488	1.113	1.679
	Low Face-Threat		
	Informative	**Hearer FM**	**Speaker FM**
Count	117	71	48
Expected Count	103.56	76.68	55.76
Adjusted Std. Residual	2.488	-1.113	-1.679

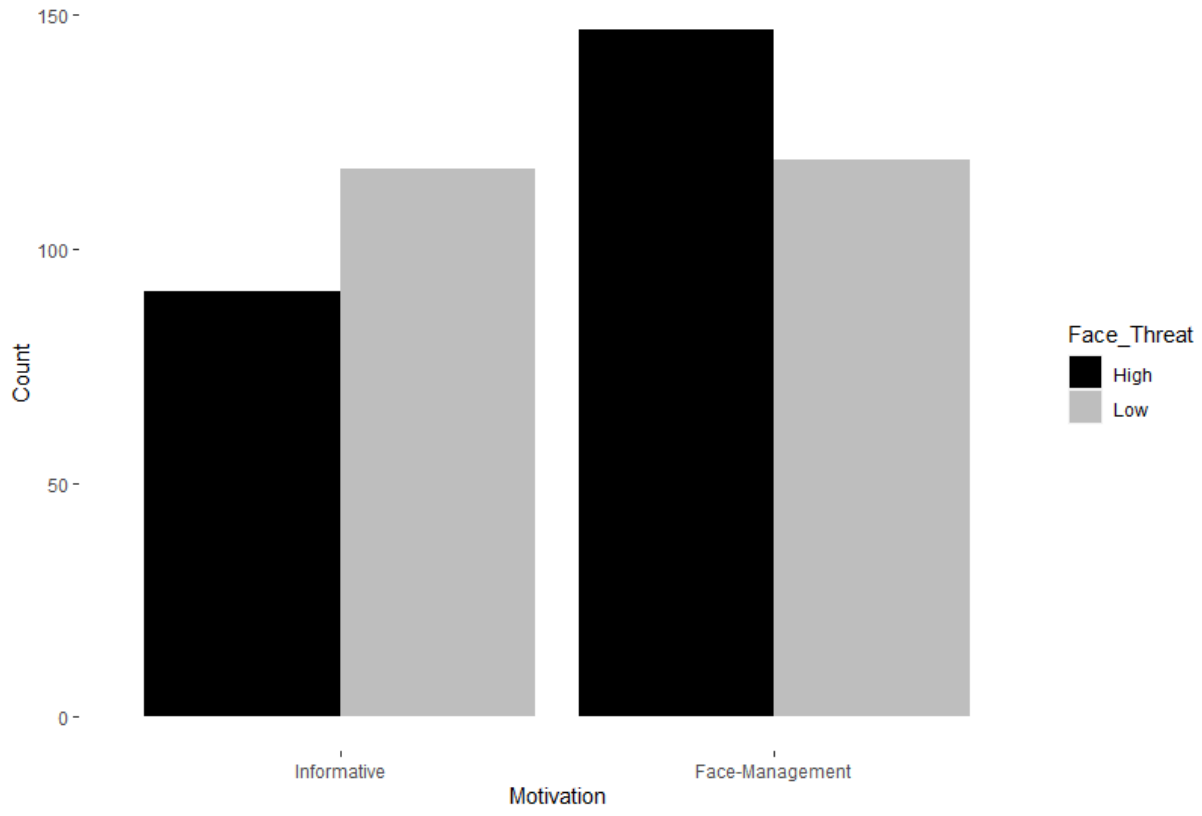
We found a significant overall association between Face-Threat and Motivation $\chi^2(2, N = 474) = 6.462$, $p = .040$, $V = .12$. The proportion of participants who were motivated by hearer face-management and speaker face-management did not differ significantly between high and low face-threat conditions ($\tilde{r}_{ij} = 1.113$ and $\tilde{r}_{ij} = 1.679$ respectively). However, there was a significant difference in the proportion of participants who were motivated by informativeness ($\tilde{r}_{ij} = 2.488$). Participants were motivated to be informative less often under high face-threat (38.2%) compared to when under low face-threat (49.6%).

Motivation (Exploratory, Not Pre-Registered)

Whilst non-significant, for both speaker and hearer face-management motivations in the original analysis, there was a higher proportion of participants under high face-threat (i.e., the same direction). We therefore considered it valuable to combine both of these motivations into a more generalised 'Face-Management' motivation, and test (via a 2x2 chi-square) whether the proportion of participants motivated by general face-management strategies differed as a function of face-threat. The findings are summarised in Figure 10.

Figure 10

Bar chart showing the association between Face-Threat (high, low) and Motivation (Informativeness, Face-Management). The y-axis represents the raw number of participants who chose each motivation, as a function of whether they were under high or low face-threat.



We again found a significant association between Face-Threat and Motivation $\chi^2(1, N = 474) = 6.189$, $p = .013$, $\varphi = .10$. Additionally though, the proportion of participants

motivated by face management differed by face-threat ($\tilde{r}_{ij}$ = 2.488). A significantly larger proportion of participants chose their quantifier based on face-management intentions under high face-threat (61.8%) compared to low face-threat (50.4%).

Association between Face-Threat and Quantifier, as a function of Motivation

Our final sample (N = 474) was subsetted into three samples, (1) those who were motivated to be informative (n = 208), (2) those who were motivated by hearer face-management (n = 154), and those who were motivated by speaker face-management (n = 112). For each of these samples, we observed the association between Face-Threat and Quantifier choice. Non-significant associations were found for Informativeness (χ^2(3, N = 208) = 3.461, p = .326, V = .13), Hearer FM (χ^2(3, N = 154) = 1.937, p = .586, V = .11) and Speaker FM (χ^2(3, N = 112) = .833, p = .841, V = .09). It should be noted here that in all of these tests, a maximum of 5 out of 8 cells reached the minimum expected frequency, suggesting insufficient power to draw appropriate conclusions.

5.1.6. Discussion

The experiment reported here builds on the design and findings of Sirota & Juanchich (2015). The authors of this paper discovered that when communicating uncertain negative information, speakers will alter the communicated probability of that information when their intentions are to achieve face-management, rather than to be informative. These findings served as the first direct empirical evidence that speakers tasked with communicating potentially upsetting information will (1) intend to deliver the information whilst also intending to manage face, and (2) in the pursuit of managing face, will deliver information that is different to the actual information they had in mind.

The results of this experiment extend these findings. We evidence firstly that in the context of 'breaking bad news', speakers will alter the probability communicated by proportional quantifiers (e.g., some, many), and they will do so as a function of how face-threatening the news is. These findings align with other experiments of uncertain language production (Holtgraves & Perdew, 2016). Secondly we evidence that when communicating upsetting health information, the extent to which a person will be motivated to engage in face-management strategies is dependent on precisely *how* threatening the news is considered to be. In short, an increase in the 'threat' of a situation is associated with a greater desire to manage the threat, as well as an increased tendency to downplay the likelihood of the threat occurring.

In line with our predictions, our manipulation of face-threat altered the distribution of quantifier choice within our data. Specifically, when communicating an uncertain medical opinion directly to the patient themselves (rather than to the patient's family doctor), a significantly larger proportion of participants chose to qualify the statement with '*some patients*', whilst simultaneously, a significantly lower proportion of participants chose to qualify the statement with '*most patients*'. The direction of this finding aligns with previous research. When being told upsetting information, hearers expect that the communicated probability of this information has been artificially 'deflated' by the speaker in an attempt to soften the blow (Bonnefon & Villejoubert, 2006; Juanchich et al., 2012). Similarly, evidence from the perspective of speakers shows that upsetting information is tactfully communicated by expressing a lower probability (Holtgraves & Perdew, 2016; Juanchich & Sirota, 2013; Sirota & Juanchich, 2015). Our data here supports these findings, showing that when face-threat was particularly high, participants chose the lower proportion quantifier ('*some patients*') more often and chose the higher proportion quantifier ('*most patients*') less often. Importantly, we evidence that this pattern of results (observed in the production of

probabilistic terms such as ‘quite likely’), extends also to the production of proportional quantifiers like ‘some’ and ‘many’.

Also in line with our predictions, our manipulation of face-threat significantly altered the distribution of participants’ motivations when choosing a quantifier. In our primary (pre-registered) analysis, a significantly smaller proportion of participants indicated that they intended to be informative when communicating an uncertain medical opinion directly to a patient (as opposed to the patient’s family doctor). This suggests in the first instance that when a situation is particularly face-threatening, a speaker’s communicative intentions will differ from that of a purely accurate information exchange (informativeness). Surprisingly, in the same analysis the occurrence of hearer face-management (tact) and speaker face-management (caution) intentions did not differ significantly between high and low face-threat conditions. In a follow-up (not pre-registered) analysis we combined both *caution* and *tact* motivation options into one, more generalised ‘face-management’ motivation. When analysing the data in this way we discovered that under high face-threat, a significantly larger proportion of participants were motivated by an intention to manage face, rather than to be informative. In agreement with previous research (Juanchich & Sirota, 2013; Sirota & Juanchich, 2015), our results, when combined, suggest that to deal with upsetting information, speakers intend to adopt general face-management strategies (rather than informativeness) that attends to either their own self-image or the self-image of their communicative partner. Crucially though, we also extend the current understanding of this effect, by evidencing that it is dependent on the threat level ‘extremity’. This could suggest that a speaker matches their intentions of engaging in face-work to their perception of how much face-work is required given the circumstances.

Finally, we were unable to evidence that face-threat is associated with communicated probability as a function of the speaker’s specific intention. It is possible therefore, that

motivations do not map neatly onto the probability communicated by a person, based solely on those motivations. However, it is more likely, based on the fact that we were unable to control for sample size per condition (i.e., samples were dictated by response selection rates rather than controlled allocation), that our three tests were insufficiently powered to detect an effect.

The experiment reported here has merit because it is one of a small number of studies assessing the production of quantifying terms within a health context. Speakers tend to prefer communicating uncertainty in words rather than numbers (Juanchich & Sirota, 2020b), and the use of proportional quantifiers like 'a few' or 'some' are among the most popular methods of achieving this by health professionals (Juanchich & Sirota, 2020a). We have shown that quantifiers can, in fact, be used to downplay or soften bad news, in the same way that other linguistic devices can, such as uncertainty terms (e.g., maybe), probabilistic statements (e.g., somewhat unlikely) and indirectness (Bonnefon et al., 2011b; Bonnefon & Villejoubert, 2006; Holtgraves & Perdew, 2016; Pinker et al., 2008; Sirota & Juanchich, 2015).

One methodological limitation to consider is that whilst the order of our quantifier scale was preserved (a few, some, many, most), none of our quantifiers adequately represented an even probability. Three-quarters of our participants considered '*some patients*' to represent a proportion on or below 50%, and more than three-quarters of our participants considered '*many patients*' to represent a proportion greater than 50%. Unlike Sirota and Juanchich (2015) then, where participants could 'anchor' their responses on the reliable middle point of the scale ('evenly probable'), the response options in this study effectively force participants (generally speaking) to go either above or below their genuine level of certainty. This is especially the case given that 92.8% of participants chose either 'some' or 'many' to communicate the news, and it could suggest that the associations observed for our face-threat manipulation may partly be explained by the limited options that represent perfect uncertainty

(50% sure). Future investigations should aim to test the reported effects across a range of hypothetical certainties to observe whether the findings generalise across varying degrees of uncertainty.

Altogether, our findings here provide further support for the postulate of politeness theory, that speakers adopting face-management intentions will alter the communicated likelihood of a negative event in order to caveat or 'cushion' bad news. Here we show (in a health context) that this is still achieved by speakers when proportional quantifiers (e.g., some, many) are the chosen form of communication. We also show that an increase in the level of face-threat within a situation is associated with less informative (and more face-managing) motivations for speech production. Considering that quantifiers do not refer to a solid (or fixed) likelihood (Holtgraves, 2014), there is a potential for misunderstandings between doctor and patient when used to communicate negative health risk. Quantifiers should therefore be used with caution, especially when communicating uncertainty under particularly high-stakes circumstances.

Acknowledgements: We thank Isabel Dunkerley who worked with us as a research assistant to assemble this study prior to data collection.

Author Contribution Statement

Study Conception: Harry Clelland, Matthew Haigh, Isabel Dunkerley

Study Design: Harry Clelland, Matthew Haigh

Development of Materials: Harry Clelland, Isabel Dunkerley

Pre-Registration: Harry Clelland, Isabel Dunkerley

Data Collection: Harry Clelland

Data Analysis: Harry Clelland

First Draft: Harry Clelland

Review and Editing: Harry Clelland, Matthew Haigh

5.2. "*A few people suffer deafness*" - Use of a Quantifier to Gauge a Severe Side Effect is Considered Tactful, but does not Increase Perceived Likelihood (Exp. 4)

5.2.1. Abstract

Research shows that the meaning of proportional quantifiers (e.g., *some* people, *many* patients) changes based on context. This is particularly important in the communication of health risk, because situational differences may cause discrepancies between the degree of risk that is intended by a health professional and the risk that is actually interpreted by a patient. This experiment sought to understand whether the severity of a side effect (i.e., severe vs. non-severe) could influence the interpretation of a risk statement qualified with a proportional quantifier (e.g., *some* people experience X). We examined four quantifiers (a few, some, many, most) and manipulated the severity of a potential side effect (severe, non-severe). Participants estimated the likelihood of experiencing the side effect, and selected what they believed to be the doctor's primary motivation for using the quantifier. We found that when used to qualify a severe side effect, participants more frequently considered the use of a quantifier (e.g., *a few*) to be a tactic for managing face (i.e., for softening the bad news), rather than an expression of genuine uncertainty. However, outcome severity did not affect perceived likelihood of the side effect occurring. Proportional quantifiers may be preferable to other non-referential terms (e.g., possibly, maybe) when communicating uncertain health risk to patients, because the likelihood that they signal may not be exaggerated under particularly threatening circumstances.

Keywords: Quantifiers, Severity, Bias, Health Communication, Uncertainty, Facework, Likelihood Perception, Experimental Pragmatics

5.2.2. Introduction

Quantifying terms like *some*, *many* and *most* refer to inexact proportions. They offer a certain amount of specificity, because they cover a limited range of potential values (Szabolcsi, 2010). For example, when stating that '*many* people make a full recovery', the term *'many'* carries a theoretical minimum and maximum proportion of people that is being referred to. However, quantifiers are also ambiguous because they represent a highly variable range of exact proportions. For example, a person who is told that '*some* people suffer complications' could interpret '*some*' to mean 5% of people, or equally to mean 95% of people, since most people infer this to mean '*some*, but not *all*' (Katsos, 2008). A person could even take '*some*' to mean 100% of people, if they were to infer '*some*, and possibly *all*' (Noveck & Posada, 2003).

Studying the interpretation of quantifiers is important because they possess a number of intriguing linguistic features. First, the proportion that they denote is largely dependent on context (Bonnefon et al., 2009; Degen & Tanenhaus, 2015; Katsos, 2008; Keenan & Stavi, 1986; Moxey & Sanford, 1986; Pezzelle et al., 2018b; Solt, 2016; Westerståhl, 1984). Stating that '*many* nurses experienced burnout this year' compared to '*many* security guards experienced burnout this year', will carry different interpretations about exactly how many workers experienced burnout, due to differences in how both professions are perceived. Second, the logical strength (or 'informativeness') of a quantifier (e.g., *some* is less informative than *all*) will impact inferences made. For example, a person hearing that '*some* people make a full recovery' will infer that '*not all* people make a full recovery' (Horn, 1984; Katsos, 2008). Third, quantifiers that express roughly the same quantity can induce different perceptions attached to that quantity (Moxey & Sanford, 2000; Nouwen, 2010). Consider a doctor who is recommending rehabilitation treatment to a patient. Stating that 'few patients

relapse after one year' will be perceived differently to stating that 'a few patients relapse after one year', even though both refer to the same proportion. This is because the former 'few patients relapse' draws attention to the majority of patients that do not relapse, whereas the latter 'a few patients relapse' draws attention to the minority of patients that do relapse. This inherent context-sensitivity becomes crucial to understand when considering the communication of health information, primarily because quantifiers are non-referential (i.e., they do not express a stable proportion; Szabolcsi, 2010), and are therefore vulnerable to ambiguity and misunderstandings (Bonnefon et al., 2009; Holtgraves, 2014; Lee & Pinker, 2010; Pinker et al., 2008).

One particular use for quantifiers is to communicate the probability of an event. Juanchich and Sirota (2020a) found that family physicians most often prefer to communicate side effect risk using verbal quantifiers rather than numbers (e.g., *a few patients experience dry mouth*, p. 14). The focus of this study is on understanding how 'proportional' quantifiers (e.g., *few, some*) are interpreted. Evidence suggests that these are the most difficult type to process behind logical (e.g., *all*), numerical (e.g., *at least*) and parity (e.g., *an even number of*) quantifiers (Szymanik & Zajenkowski, 2010; Zajenkowski & Szymanik, 2013; Zajenkowski et al., 2014). Furthermore, the choices made by recipients on the basis of a verbal quantifier is largely context-dependent (Liu et al., 2020), resulting in an impaired ability to make decisions (Liu et al., 2021). This evidence suggests that when side effect risk is expressed with a proportional quantifier, this may cause confusion over the likelihood of experiencing the side effect, whilst also impairing a patient's ability to make health-based decisions as a consequence.

Studies to date have evidenced a number of pragmatic factors that influence the perception of 'some' (Bonnefon et al., 2015; Degen & Tanenhaus, 2015; Loy, Rohde &

Corley, 2019; Mazzarella et al., 2018). More specifically, research has consistently demonstrated that when 'some' is used in a face-threatening context (i.e., when the information being communicated threatens a person's social identity), interpretations are influenced by the perceived motivations of the speaker (Bonnefon et al., 2011a; Bonnefon & Villejoubert, 2006). For example, consider a patient who is told by their doctor that '*some* people experience side effects'. This statement implies that 'possibly all' people experience side effects, and given that the statement threatens the hearer's face, the inference that 'some, but not all' people experience side effects becomes a less available interpretation (Bonnefon et al., 2009). In other words, the patient considers the use of '*some*' to be a face-management tactic used to soften the threat of a serious act.

Beyond the quantifier 'some', there is growing evidence demonstrating that a severe outcome affects the interpretation of other non-referential terms. Bonnefon and Villejoubert (2006) presented participants with a medical scenario, in which they are told by their doctor that they will '*possibly*' experience a mild or severe medical condition (insomnia vs. deafness). Their findings reflected the severity bias (Fischer & Jungermann, 1996; Weber & Hilton, 1990), whereby participants considered a relatively severe outcome such as '*possible* deafness' to be more likely than a relatively less severe outcome, such as '*possible* insomnia' . By extension, participants considered the non-referential qualifying term '*possible*' to be a face-management tactic when the medical condition was severe. The findings highlight the importance of perceived politeness motivations in medical communication, because a misinterpretation of a probabilistic phrase can cause a discrepancy between the likelihood that the doctor intends to communicate, and the likelihood that is actually understood by the patient.

The current study aims to understand whether the effects of the severity bias observed on the use of uncertainty terms (e.g., possibly, maybe), extends to proportional quantifiers (e.g., a few, many). We believe this to be a necessary development firstly because severity can cause a patient to overestimate their likelihood of catching a disease or experiencing a side effect (Cox, 2019). Couple this with the preference for family physicians to report side effect risk with verbal quantifiers over numbers (e.g., '*a few* patients experience fevers'; Juanchich & Sirota, 2020a), and it becomes important to establish whether a severe event qualified with a proportional quantifier can firstly elevate the interpreted likelihood of the event, as well as establish if the inflation of likelihood perception can be explained, at-least in part by interpretations of politeness.

Quantifying terms are particularly intriguing within this context because, unlike uncertainty terms, they offer a level of specificity. That is, whilst they do not refer to a stable proportion, quantifiers are mentally arranged on an ordered (but non-linear) scale (Pezzelle et al., 2018a). In other words, the mental representation of quantifiers is such that *few* represents a smaller proportion than *some*, which represents a smaller proportion than *many*, and so on. Research has already established that the perception of *some* is altered under face-threat (Bonnefon et al., 2009). Bonnefon et al. (2009) asked participants to imagine that they had read a poem for an audience. They discovered that when *some* was used in a context that could boost the face of the recipient, being told that '*some* people loved your poem' caused 17% of participants to reason that maybe everyone loved their poem. By contrast, when some was used in a context that could threaten the face of the recipient, being told that '*some* people hated your poem' caused 42% of participants to reason that maybe everyone hated their poem. Applied to health communication, it may be the case that a person will consider the proportion denoted by '*some* people' to be higher when facing a relatively severe negative health

outcome (e.g., '*some* people experience insomnia' vs. '*some* people experience deafness'). This study aims to test this empirically. Moreover, we aim to establish whether the interpretation of other relevant proportional quantifiers (a few, many, most) are also sensitive to the severity bias.

5.2.3. The Present Research

Developing research into the interpretation of non-referential terms (e.g., *possibly*) under face-threatening contexts, this study utilises the pragmatic complexities of health communication to establish whether the severity bias, whereby people consider more severe events to be more likely, influences the likelihood perception of health outcomes when they are qualified with proportional quantifiers (e.g., '*some* people experience side effects'). Additionally, we aimed to establish whether the use of a proportional quantifier to qualify a severe health outcome would be interpreted as a politeness marker. In other words, is the quantifier considered a way to tactfully deliver the bad news, rather than a genuine expression of proportion perception held by the speaker.

We asked participants to consider a fictional scenario where they imagine themselves as a healthcare patient. They are offered a new drug by their doctor that will effectively treat a disease that they have recently been diagnosed with, however, the treatment comes with a potential side effect. Severity of the side effect was manipulated in our scenario by creating two versions. In one version, participants are told that the potential side effect is insomnia. This is the low severity side effect. In another version, participants are told that the potential side effect is deafness. This is the high severity side effect. These medical conditions are the same as those used by Bonnefon and Villejoubert (2006) to manipulate severity. In all scenarios, the possibility of getting the side effect is qualified by the doctor, using a proportional quantifier. The specific quantifier seen by the participant was either '*a few*',

'*some*', '*many*' or '*most*'. For example, in the high severity scenario, participants may have been told that '*some* people suffer deafness', or in the low severity scenario, told that '*many* people suffer insomnia'.

After reading the scenario, participants were first asked what they believed the likelihood of getting the side effect to be. After this, participants were asked to judge whether the quantifier (either a few, some, many or most) was used by the doctor because (1) they did not know exactly how many people experience the side effect, or (2) they wished to announce the news tactfully.

For our likelihood measure, we predicted that there would be a main effect of severity, whereby across all quantifiers, the perceived likelihood of experiencing deafness (high severity) would be higher than the perceived likelihood of experiencing insomnia (low severity). We also predicted that there would be an interaction between severity and quantifier. Specifically, we expected that the severity bias would have the largest impact on quantifiers that represent lower proportions. That is, the difference in likelihood perception between high and low severity, will be larger for the difference between '*a few* people suffer deafness' (vs. insomnia) than it is for the difference between '*most* people suffer deafness' (vs. insomnia). The reason for our prediction is that the latter is less likely to be perceived as a face management tactic. Bonnefon and Villejoubert (2006) evidenced that '*possible* deafness' was considered to be more likely than '*possible* insomnia', and that this discrepancy was explained by participants' belief that the qualifying term *possible* was used to manage social threat in the high severity condition. In other words, participants were able to detect that the sensitivity of the situation would need to be managed by the speaker, and subsequently concluded that the reason for an under-informative qualifier (i.e., *possibly*) was to mitigate the threat posed by the speaker's words. In the context of our medical scenario, we predicted that a participant who

reads '*most* people experience deafness' is unlikely to see the use of *most* as a tactic for managing face, because compared to a lower proportion quantifier (e.g., *a few* people suffer deafness), the statement does not appear to 'downplay' the severity of the prognosis. As such, we predicted that lower proportion quantifiers (e.g., *some* people) will be judged as a face-management tactic more frequently when the side effect is severe, causing a greater discrepancy between the likelihood perception of low and high severity side effects, than when those same side effects are qualified with higher proportion quantifiers (e.g., *many* people).

5.2.4. Method

This online study was pre-registered prior to data collection (see osf.io/gzcra). Raw data, analysis script and materials can be found on the Open Science Framework (see osf.io/wat4g/). The aim of this online experiment was to test whether severity of an outcome would impact judgements of medical likelihood (i.e., replicating the 'severity effect'), across a range of different proportional quantifiers (e.g., *some, many*). We also aimed to test whether perceived motivations for the use of a proportional quantifier would differ between high and low severity outcomes (either genuine uncertainty or a politeness marker). Participants imagined themselves as a medical patient faced with the possibility of experiencing a side effect. The doctor who is communicating this risk, attributes the prevalence of the side effect to either 'a few people', 'some people', 'many people' or 'most people'. The actual side effect faced is either low in relative severity (insomnia) or high in relative severity (deafness). Based on these scenarios, we measured participants' perceptions of likelihood, as well as their perception of the doctor's primary motive for using the quantifier.

Design

We utilised a two-factor (4x2) independent groups design. Our first factor was the Quantifier used by the doctor, with four levels ('a few', 'some', 'many', 'most'). Our second factor was Severity of the side effect, with two levels (low severity = insomnia, high severity = deafness). Participants were randomly allocated to one of eight experimental conditions (see Table 5).

Table 5

Eight experimental conditions in our 4x2 design.

Severity	**'a few' people**	**'some' people**	**'many' people**	**'most' people**
Low Severity (insomnia)	a few/Low	some/Low	many/Low	most/Low
High Severity (deafness)	a few/High	some/High	many/High	most/High

Our study had two dependent measures. The first was a participants' estimate of likelihood (i.e., perceived risk of getting the side effect). The second was perceived motivation. This measure asked participants to determine the doctors' primary Motivation for using the quantifier (either genuine uncertainty or politeness).

Participants

Our minimum sample was determined based on the least powerful set of tests (χ^2 tests, df = 3). A prospective power analysis was conducted using the 'pwr' R package (Champely, 2020), based on two post hoc tests (see our pre-registration for details of all planned chi-square tests; osf.io/gzcra). The minimum value to detect a medium effect size 80% of the time (using α = .025) was 143.4. Given that each test uses half of the sample, we required a minimum sample of 287 (143.4 x 2 = 286.8).

Participants were recruited via opportunity sampling, advertised online through social media and university mailing lists. We collected an initial sample of 716 consenting volunteers. We removed those who did not complete the study (n = 42) and those who declared a non-serious response (n = 1). Our final sample comprised 673 English-speaking adults (150 men, 504 women, 12 non-binary, 7 self-identify) aged between 18 and 71. The number of participants allocated to each condition ranged between 80 and 87.

Materials

We created a single experimental vignette for this study (see Figure 11). The vignette places participants in a hypothetical scenario, in which they imagine themselves as a health patient recently diagnosed with Muscular Dystrophy. They are offered a new treatment by their doctor, which has a potential negative side effect. Having asked about their risk, they are told that either 'a few people', 'some people', 'many people' or 'most people' experience the side effect. The severity of the potential side effect was either relatively low (insomnia) or relatively high (deafness). Considering the scenario, participants were asked to estimate their likelihood of experiencing the side effect (out of 100%). They were also asked to judge whether the doctors' primary motivation for using the quantifier (e.g., some) was either genuine uncertainty, or an expectation of politeness.

Figure 11

Example vignette. Information in italics was given to all participants and was sourced from the NHS website (UK).

The muscular dystrophies (MD) are a group of inherited genetic conditions that gradually cause the muscles to weaken, leading to an increasing level of disability. MD is a progressive condition, which means it gets worse over time. It often begins by affecting a particular group of muscles, before affecting the muscles more widely. Some types of MD eventually affect the heart or the muscles used for breathing, at which point the condition becomes life-threatening. There is no cure for MD, but treatment can help to manage many of the symptoms.

Imagine that you are a patient at a local medical practice. Last week you were diagnosed with early onset Muscular Dystrophy, and you are at a consultation with your doctor. Your doctor offers you a new medication to treat your condition. The new drug is effective at stopping the disease from getting worse.

You ask your doctor about potential side effects, and he tells you that "*(a few/some/many/most) people suffer (insomnia/deafness) after taking this medication*".

Quantifier Manipulation

Four scenario versions were created to test four levels of Quantifier. In the first, the doctor states that '*a few*' people experience the side effect. In the second, third and fourth versions, the risk is attributed to either '*some people*', '*many people*' or '*most people*' respectively.

We chose to examine the quantifiers used by Pezzelle et al. (2018a), whose experiments probed the mental representation of nine quantifying terms. However, in this experiment we are interested only in the perception of *proportional* quantifiers (e.g., some, many; Szymanik & Zajenkowski, 2010; Zajenkowski & Szymanik, 2013; Zajenkowski et al., 2014). These are the hardest type of quantifier to process, yet are used most frequently by

family physicians to communicate risk (Juanchich & Sirota, 2020a). We therefore reduced Pezzelle's quantifiers list by eliminating four *logical* quantifiers (none, almost none, all, almost all) and one *numerical* quantifier (the smaller part).

Severity Manipulation

Two scenario versions were created to operationalise Severity within our vignette. In one version, participants are faced with the possibility of experiencing insomnia (low severity) as a side effect of the new medication. In the other version, they are faced with the possibility of deafness (high severity). The decision to opt for insomnia/deafness specifically, is in replication of Bonnefon and Villejoubert (2006) who tested 'possible insomnia' vs. 'possible deafness'. They report that of their final sample (N = 941), 86.1% considered deafness to be more severe than insomnia. Based on their data, we considered the extremity of our severity manipulation to be sufficient.

Measures

Likelihood Estimate

After reading the scenario, participants were asked '*In your opinion, if you accept the new medication, what is your likelihood of getting this side effect?*'. Subjective ratings of likelihood were provided as a percentage, on a slider scale ranging from 0 to 100. The scale was labelled at 0, 25, 50, 75 and 100.

Perceived Motivation

Perceived motivation of the doctor was obtained through a dichotomous measure. On a new page, participants were reminded of their doctors' statement:

'*Your doctor said:*

(a few/some/many/most) people suffer (insomnia/deafness) after taking this medication'

They were then asked: '*In your opinion, what is the main reason why the doctor said '(a few/some/many/most)' people?*'. Participants had to decide whether the doctors' primary motivation for using the quantifier was either genuine uncertainty (a) '*He was not sure if you would get the side effect*' or a politeness marker (b) '*He wished to announce the news tactfully*'. For those participants unfamiliar with 'tact', an oxford dictionary definition was provided via link.

Procedure

This online study was run using Qualtrics survey platform (Snow & Mann, 2013). The study was accessed via an anonymous link. After providing basic demographic information, participants saw a scenario preface that reads as the following:

"In this short survey you will be given a hypothetical scenario. You will be asked to imagine that you are a healthcare patient. Please read this scenario, then answer the two questions that follow based on your judgement of the situation."

Once briefed, participants were randomly assigned to one of 8 scenarios (4x2 conditions; see Table 1). After reading the scenario, participants estimated their likelihood of experiencing the given side effect. Then on a new page, they indicated their perception of the doctors' motivation for using the given quantifier. Before submitting the survey, participants were asked whether they took the task seriously (Aust et al., 2013). The end-of-survey message provided debrief information. Median completion time for the full survey was 2.1 minutes. The experiment can be replicated exactly using a .qsf file for the survey, available via the OSF (see osf.io/evjby/).

5.2.5. Results

The following analysis was pre-registered prior to data collection. Our pre-registration, raw data and analysis script can be accessed on the OSF (see osf.io/wat4g/).

A response was required for each question, so there was no missing data and no data points were excluded. Analyses were conducted using R version 4.1.0 (R Core Team, 2021) and RStudio version 1.4.1717 (RStudio Team, 2021).

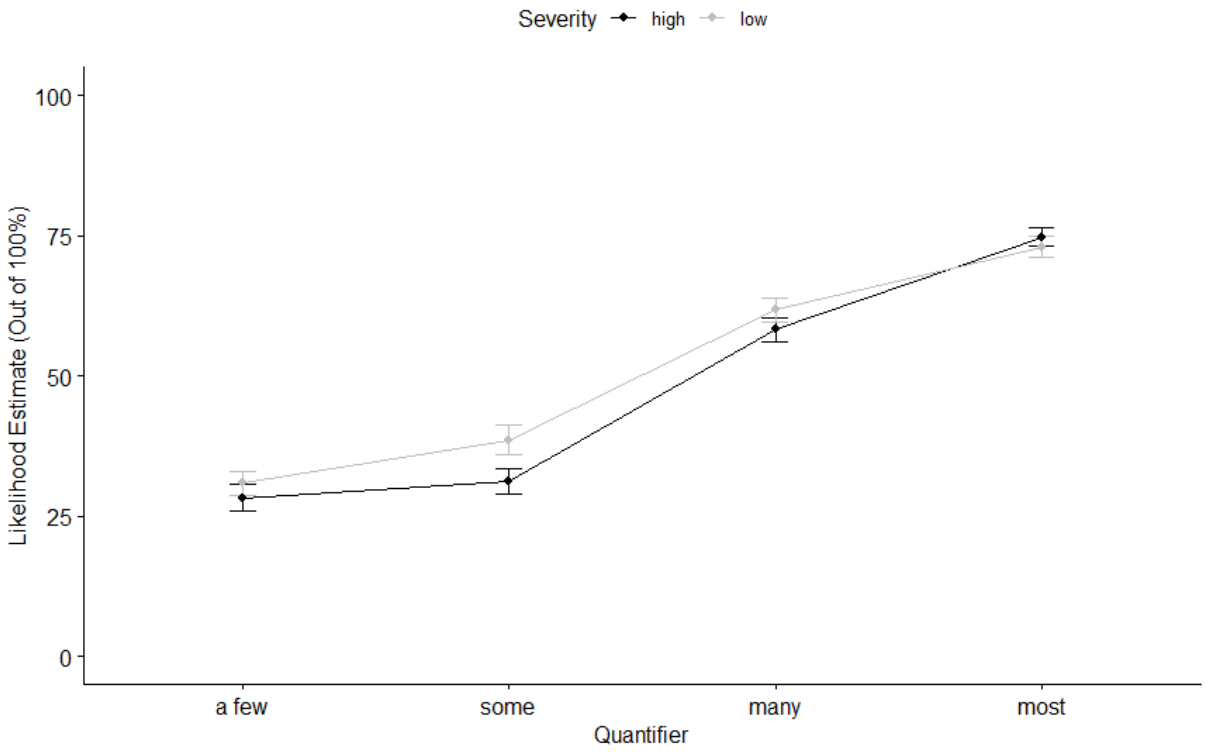
To analyse our Likelihood Estimate measure, we ran a two-factor (4x2) independent groups ANOVA using the 'aov' base R function. Cohen's f (*f*) is reported as a measure of effect size. Our perceived motivation measure was analysed using Chi-square (χ^2) tests of independence, using the 'gmodels' package (Warnes et al., 2018). We also calculated Adjusted Standardised (Pearson) Residuals ($\tilde{r}_{ij}$) alongside each test, to determine whether cell counts deviated from their expected counts. Cells that exceed the absolute value = +/-1.96 indicate a significant deviation from the expected count. Phi (φ) and Cramer's V (V) are reported as measures of effect size.

Likelihood Estimate

After reading the scenario under one of our eight experimental conditions, we measured participants' perceptions of how likely they were to experience the side effect. Findings are displayed in Figure 12.

Figure 12

Line graph showing mean likelihood estimates across four proportional quantifiers at high and low severity. Error bars represent standard error.



Unsurprisingly, we found a main effect of Quantifier $F(3, 665) = 190.902$, $p < .001$, $f = 0.92$. However, we did not find a significant main effect of Severity $F(1, 665) = 3.808$, $p = .051$, $f = 0.06$ or a significant interaction between Quantifier and Severity $F(3, 665) = 1.528$, $p = .206$, $f = 0.06$.

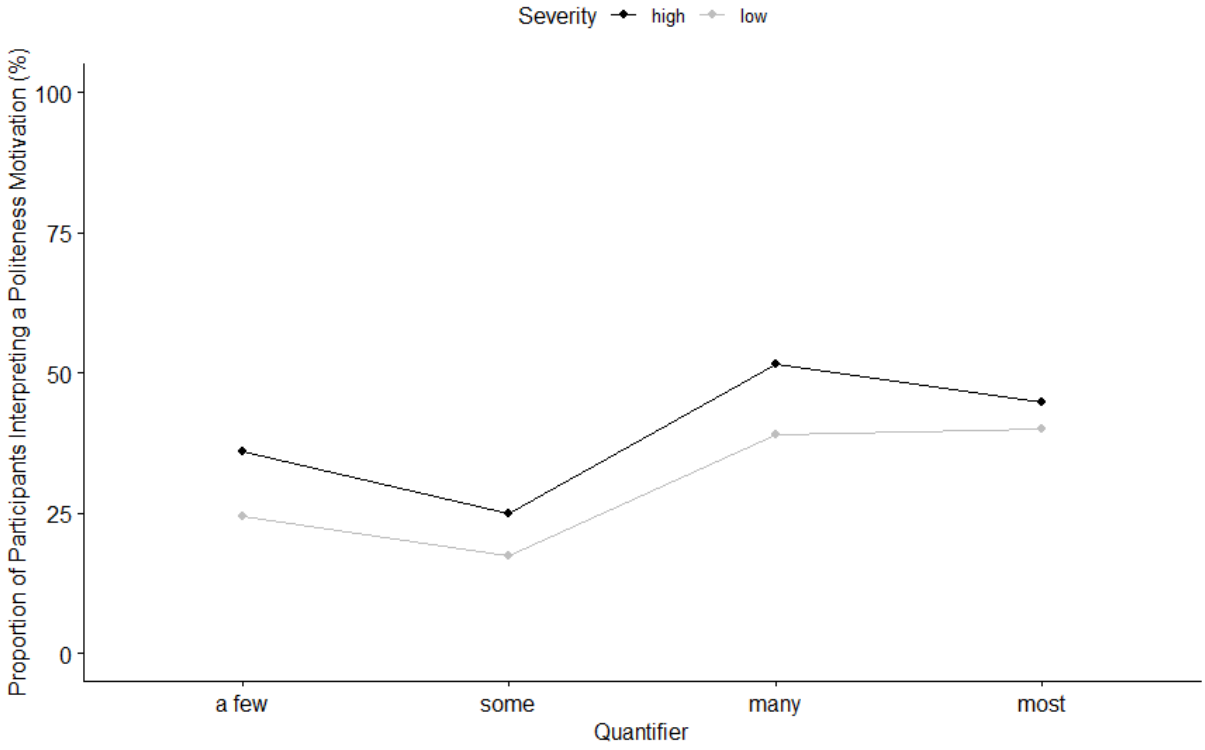
Perceived Motivation

After reading the scenario under one of our eight experimental conditions, our participants indicated what they believed to be the doctor's primary motivation for using that specific quantifier (e.g., *some, many etc.*). As described above, participants were asked to choose one of two options. Option 1: the primary motivation was genuine uncertainty (i.e. the

doctor was not sure if they would experience the side effect) Option 2 the primary motivation was a politeness tactic to manage face (i.e., the doctor wishes to announce the bad news tactfully). Findings are summarised in Figure 13.

Figure 13

Line graph showing the proportion (%) of participants who interpreted the use of a quantifier as a politeness strategy over genuine uncertainty, as a function of quantifier (a few, some, many, most) and severity (high, low).



Quantifier

There was a significant association between Quantifier and perceived Motivation $\chi^2(3, N = 673) = 25.569$, $p < .001$, V = .20. Broken down by individual quantifier, 30.4% of participants considered the use of '*a few* people' to be a politeness marker. 22.3% of participants considered the use of '*some* people' to be a politeness marker. 45.4% of participants considered the use of '*many* people' to be a politeness marker. Finally, 42.4% of

participants considered the use of '*most* people' to be a politeness marker. The significant association between Quantifier and Motivation occurred both when severity was Low $\chi^2(3, N = 331) = 12.455$, $p = .006$, V = .19, and when severity was high $\chi^2(3, N = 342) = 14.228$, $p = .003$, V = .20.

Severity

There was a significant association between Severity and perceived Motivation $\chi^2(1, N = 673) = 5.527$, $p = .019$, $\varphi = .08$. When severity was Low (insomnia), 30.8% of participants considered the use of a quantifier to be a politeness marker. This was a significantly lower proportion than the expected cell count ($\tilde{r}_{ij} = -2.351$). When severity was High (deafness), 39.5% of participants considered the use of a quantifier to be a politeness marker. This was a significantly higher proportion than the expected cell count ($\tilde{r}_{ij} = 2.351$).

5.2.6. Discussion

The aim of this experiment was to extend the current understanding of the severity bias and its effects on perceptions of likelihood. That is when severe events are judged to be more likely to occur than non-severe events (Fischer & Jungermann, 1996; Weber & Hilton, 1990). When a statement of medical probability is qualified with a non-referential term (e.g., you will *possibly* experience X), a person will consider the likelihood of this occurring to be more likely when the outcome is severe (Bonnefon & Villejoubert, 2006). This bias is evidenced as true for the quantifier '*some*' (Bonnefon et al., 2009) but not for other proportional quantifiers (e.g., *many*). The aim of this study was therefore to establish if the consequences of the severity bias extend to other quantifiers of this category. Contrary to our predictions, on our likelihood measure we found no effect of severity and no interaction between quantifier and severity. In other words, the severity of the potential outcome (side-effect) did not influence

how likely that outcome was perceived to be, irrespective of the quantifier used to qualify the statement. However, in line with our predictions, we found that perceived motivations for the use of a quantifier were sensitive to the severity bias. When the outcome was high in severity (deafness), participants reliably considered the use of a quantifier to be a tactic for managing face rather than a genuine expression of uncertainty, compared to when the outcome was low in severity (insomnia).

Our findings suggest that within the context of medical risk communication, the likelihood that is expressed by proportional quantifiers does not change depending on the severity of the prognosis. These findings differ from similar experiments on uncertainty terms (e.g., possibly, maybe). Bonnefon and Villejoubert (2006) found that the likelihood of 'possible insomnia' was considered to be less probable than the likelihood of 'possible deafness'. Here, the data suggest that this is not true of proportional quantifiers, such that '*a few* people suffer insomnia' invokes the same probabilistic interpretation as '*a few* people suffer deafness'. This was particularly surprising for the comparison between '*some* people suffer insomnia' and '*some* people suffer deafness', because the interpretation of this quantifier is known to differ under face-threatening circumstances. Bonnefon et al. (2009) demonstrated that when interpreting *some*, participants made the scalar inference '*some,* and possibly *all*' more often when the statement being interpreted was face-threatening. That is, a significantly larger proportion of participants would reason that '*maybe everyone* hated their poem' as opposed to '*maybe everyone* loved their poem'. Considering the experimental differences between Bonnefon's 2009 paper and the current study, perhaps it is the case that a face-threatening situation impacts the interpretation of *some*, but the severity of information within that face-threatening context does not. In other words, it may not matter *how* threatening the situation is, so long as it's threatening in the first instance. This may be good

news for the communication of health risk to patients, because whilst non-referential uncertainty terms cause a discrepancy between the likelihood intended and likelihood interpreted (Bonnefon et al., 2011b; Bonnefon & Villejoubert, 2006; Harris & Corner, 2011), the use of quantifiers, by contrast, may be minimising the potential for misunderstandings.

We found a significant association between outcome severity and interpretations of a politeness motivation, whereby participants considered the use of all four quantifiers to be a face-management tactic more often when the outcome was high in severity (deafness) compared to low in severity (insomnia). Returning to Bonnefon and Villejoubert's 2006 findings on 'possible insomnia' vs. 'possible deafness', the authors discovered that the difference in perceived likelihood observed between the two statements was explained by participants' expectations of politeness. In other words, participants facing deafness considered the use of the term '*possible*' as a face-management tactic to reduce the impact of the statement, rather than a genuine expression of uncertainty and as a result, considered themselves more likely to suffer the side effect. Our data extends these findings, evidencing that a person will make the same politeness interpretation under high severity (deafness), when the statement is qualified with a proportional quantifier (e.g., *some* people suffer deafness). What is particularly intriguing though, is that in our data, politeness expectations did not result in a change in perceptions of likelihood. This is not in keeping with existing research in this area. Sirota and Juanchich (2012b) evidenced that even exact numerical probabilities (e.g., 50%) are sensitive to politeness expectations, and that this interpretation led to varying perceptions of risk. We instead evidence here that for the interpretation of quantifiers, their use under high severity will trigger a perceived politeness motivation, but this does not then translate into an altered perception of risk. This is a finding that future research should seek to untangle.

Our experiment here is valuable firstly because we are one of a minority of studies that have sought to understand the pragmatic factors that influence the interpretation of quantifiers within a health context. Family physicians most often prefer verbal quantifiers (rather than numbers) for communicating the risk of side effects to patients (Juanchich & Sirota, 2020a), and as such the need to understand the contextual factors that may influence likelihood interpretations based on their use becomes paramount. Moreover, the scenario that we gave to participants was of particular utility in this pursuit. Health professionals are most likely to communicate risks that are associated with medications or treatments (Fagerlin et al., 2011), making our scenario (albeit hypothetical) directly applicable to commonly encountered doctor-patient interactions.

It is also important to consider the limitations of this experiment. We must consider the possibility that our severity manipulation was not strong enough to induce the hypothesised effects. Our decision to utilise 'insomnia' and 'deafness' as our low and high severity side effects respectively, was in replication of Bonnefon and Villejoubert (2006) whose participants considered deafness to be more severe than insomnia by a very large margin (>85%) when asked to choose between the two. However, the current experiment placed participants in independent groups rather than across repeated measures conditions, meaning that our participants could not 'anchor' their judgement based on a comparative interpretation. It may be the case that if participants were to rate both side effects on a continuous scale, their ratings would suggest a smaller discrepancy in perceived severity than is shown through a forced-choice measure. Future research may wish to operationalise severity by manipulating outcomes that are 'further apart' in extremity.

A further methodological limitation here is that our dichotomous measure of politeness motivation may have hindered our ability to see any true differences between lower and higher

proportion quantifiers (e.g., a few vs. many). Participants were asked what they believed the doctor's primary motivation was for using the quantifier given. Either they thought that the doctor was unsure if they would get the side effect, or they wished to announce the news tactfully. However, genuine uncertainty becomes a less sensible option when the quantifier depicts a high proportion. For example, a doctor who states that '*most* people suffer deafness' is likely to sound a lot more certain that the patient will get the side effect, than a doctor who states that '*a few* people suffer deafness'. It is possible then that participants in the '*many'* and '*most'* conditions selected politeness as the motivation, simply because genuine uncertainty was a less viable option. This could explain why we saw a higher proportion of participants selecting politeness for *many* and *most* (45%, 42%) compared to *a few* and *some* (30%, 22%).

Overall, this experiment adds to our understanding of outcome severity and its influence on perceptions of health risk. We evidence that within a face-threatening context, a person who faces a high severity outcome (compared to a low severity outcome) will consider the use of a quantifier (a few, some, many, most) to be a politeness tactic used for managing face, in the same way that uncertainty terms (e.g., possibly) are interpreted as a communicative tool for managing face. However, severity does not appear to alter the perception of health risk emitted by a quantifier. Quantifiers may be preferable over other non-referential terms in the communication of health risk, because even though they can be seen as politeness markers in high-stakes situations, the likelihood that they depict does not appear to be unduly influenced by a particularly severe outcome.

Author Contribution Statement

Study Conception: Harry Clelland, Matthew Haigh

Study Design: Harry Clelland, Matthew Haigh

Development of Materials: Harry Clelland

Pre-Registration: Harry Clelland

Data Collection: Harry Clelland

Data Analysis: Harry Clelland

First Draft: Harry Clelland

Review and Editing: Harry Clelland, Matthew Haigh

5.3. "*There is a 1 in 10 chance of catching Ebola*" – The 1-in-X Bias is Independent of Outcome Severity (Exp. 5, 6)

5.3.1. Abstract

Research has consistently demonstrated a '1-in-X' bias, whereby people consider a 1-in-X ratio (e.g., 1 in 100 chance of catching a disease) to be more likely than an equally probable N-in-X ratio (e.g., 5 in 500). It is also known that 'possibly' experiencing a severe side effect is considered more likely than 'possibly' experiencing a non-severe side effect (the severity bias). Experiment 1 (N=697) set out to establish whether severity of the outcome (e.g., severe / non-severe disease) moderates the 1-in-X bias. Experiment 2 (N=621) focused on how these factors impact health-based decision making. In Experiment 1, participants imagined facing the possibility of catching a disease. Likelihood was expressed as either a '1-in-X' ratio (e.g., 1 in 7) or a 'N-in-X*N' ratio (e.g., 10 in 70), and the severity of the disease was either low (Lyme Disease) or high (Ebola). Participants estimated disease risk and made health decisions based on this risk. We replicated both biases; '1-in-X' ratios were perceived as more probable than N-in-X*Ns ratios, while the severe outcome was perceived as more probable than the non-severe outcome. However, severity did not moderate the 1-in-X bias. The effect size of the 1-in-X bias was the same for both severe and non-severe outcomes, suggesting that the biases independently cause overestimations of health risk. On decision making, a high severity outcome resulted in a significantly larger proportion of participants taking decisions to avoid this outcome (e.g, vaccination), but ratio format did not influence the decision. Phasing out '1-in-X' ratios from the communication of medical likelihood offers a cost-effective method of improving the accuracy of health communication.

Keywords: 1-in-X, Severity, Cognitive Bias, Health Communication, Health Risk

5.3.2. Introduction

Expressing a likelihood as '1-in-X chance of Y', compared to a 'N-in-X*N chance of Y' reliably produces a bias that is consistent across cultures, educational status and outcomes (Pighin et al., 2011; Pighin et al., 2015; Sirota et al., 2014; Sirota et al., 2018a; Sirota & Juanchich, 2019). For example, imagine that a person is planning a trip to Kenya, and that on this trip they face the possibility of contracting Malaria. A person who is informed that they face a 1 in 200 chance of being infected (1-in-X), typically considers themselves to be at greater risk than a person who is informed that the chance is 5 in 1000 (N-in-X*N), despite the fact that their probabilities are identical (Pighin et al., 2011). A person may therefore change a decision about their health, such as to accept or refuse a vaccine, based simply on the format in which risk information is presented.

Recently, researchers have declared the previously established '1-in-X' effect to be a form of cognitive bias, causing an overestimation of probability relative to 'N-in-X*N' ratios (Sirota et al., 2018a). Identifying such a bias was an important development because it supported calls to retire the '1-in-X' format in the communication of health risk (Zikmund-Fisher, 2011, 2014). For instance, widespread use of the '1-in-X' ratio is potentially problematic within the UK's National Health Service, where they are overwhelmingly favoured by health professionals and frequently used in online information articles (Sirota et al., 2018b). Continuing to understand the impact of the '1-in-X' bias is particularly important, because small cognitive misrepresentations of health risk can be magnified when adopted on a large scale.

As well as ratio format, the severity of the outcome affects the perception of medical likelihood. For example, a person who is told by their doctor that they will 'possibly suffer insomnia' typically considers this outcome to be less probable than a person who is informed

of a more severe outcome (e.g., 'possibly suffer deafness'; Bonnefon & Villejoubert, 2006). This is an example of the 'severity bias'. That is, when a person judges a negative health outcome to be more likely, the more severe the prognosis. In other words, when a statement is read about the possibility of a patient contracting a disease or experiencing a negative side effect, the likelihood of this occurring is judged as higher, the more severe the condition (Fischer & Jungermann, 1996; Weber & Hilton, 1990). An important implication here is that two equal probabilities may be interpreted differently based on the severity of the information (e.g., a 1 in 5 chance of deafness may be considered more likely than a 1 in 5 chance of insomnia).

Considering the above, it is clear that there are two known biases that can influence the perception of health risk during communication. First, expressing a likelihood as a '1-in-X chance of Y' rather than a 'N-in-X*N chance of Y' influences the perceived likelihood of the occurrence of Y. Second, altering the severity of outcome Y (i.e., facing the possibility of deafness rather than insomnia) also influences the perceived likelihood of the occurrence of Y.

Advancing understanding in this area, the current work sets out to establish if the two biases interact - is the 1-in-X bias moderated by the severity of the outcome? We believe this to be a necessary development primarily because health professionals prefer '1-in-X' ratios over both 'N-in-X*N' ratios and percentages, particularly when the health outcome is negative (Sirota et al., 2018b). Given the prevalence of these ratios in medical communication, even small effects on likelihood perception are relevant to public health.

Specifically, we expect that the '1-in-X' bias will, in fact, be moderated by the severity of an outcome. We predict this for two key reasons. Firstly, people in general tend to overestimate the probability of severe outcomes (Harris & Corner, 2011). Secondly, Sirota and colleagues discuss that the '1-in-X' bias may activate this overestimation mechanism, and

encouraged other researchers to test this directly (Sirota et al., 2018a). In other words, it is possible that the mechanism causing an overestimation of health risk when the outcome is severe, is the same mechanism that is activated when exposed to a '1-in-X' ratio. Research should therefore test whether the '1-in-X' bias is *amplified* by a high severity outcome. This is because studies to date have operationalised the '1-in-X' bias in scenarios with a constant outcome, whereas health professionals are most likely to communicate the risks associated with medications or treatments. In other words, contextual situations that have highly variable outcomes.

For context, consider again a patient who is faced with the possibility of experiencing a negative side effect from medication. The true likelihood of this occurring is 14.3%. Expressing this likelihood to the patient as a 1 in 7 chance, causes a larger overestimation of risk than if the likelihood were to be expressed as a 10 in 70 chance. However, we expect that the extremity of this effect will be dependent on how severe the side effect is. For example, if the side effect is insomnia (less severe), the difference in perceived likelihood between a 1 in 7 chance of insomnia, and a 10 in 70 chance of insomnia, is likely to be less pronounced than if the side effect is deafness (more severe). In short, perhaps it is the case that in severe situations, a person is more susceptible to the '1-in-X' bias. Understanding the interplay between both sets of biases is therefore an important step forward in grasping the practical implications of medical likelihood communication.

The experiments reported below build on two recent papers establishing the clinical significance of the '1-in-X' bias. Sirota et al. (2018a) presented participants with a hypothetical scenario in which they are planning a trip abroad, and face the possibility of contracting a disease while travelling. Across five experiments, exposure to a '1-in-X' ratio resulted in a larger overestimation of probability than 'N-in-X*N' ratios. Participants in the '1-

in-X' condition also indicated on a verbal probability scale that they would be more likely to cancel their trip, evidencing that the '1-in-X' bias has the potential to impact health-based decision making.

Acting on the need to test if the '1-in-X' bias could impact actual health-based dichotomous decisions, Sirota and Juanchich (2019) gave participants a similar scenario (i.e., planning a trip to Kenya), with which they replicated the '1-in-X' bias. However, they also asked participants to imagine that they were suffering from a disease, and to indicate if they would accept treatment based on a risk of side effects. On these measures, the '1-in-X' bias made a small but significant impact on dichotomous decision making. In other words, the ratio format observed significantly impacted whether participants chose to accept or refuse a medical treatment. In this paper, we report two experiments that set out to replicate and extend these recent works.

5.3.3. Experiment 1

The aim of Experiment 1 is firstly to replicate the findings of Sirota et al. (2018a). That is to test whether '1-in-X' ratios (e.g., 1 in 3) produce a larger overestimation of health risk relative to 'N-in-X*N' ratios (e.g., 13 in 39), whilst subsequently affecting health-based decision making. Beyond replication, this experiment also aims to establish whether the severity of an outcome moderates the '1-in-X' bias.

We presented participants with a hypothetical scenario. Within this, they imagined that they were planning a trip abroad to Sierra Leone, and are told by their doctor about the possibility of contracting a disease whilst on their trip. Participants estimated their chances of contracting the disease, and indicated with a Yes or No whether they would accept a safe vaccine against the disease, were they to be offered one.

We manipulated our scenario in two ways. Ratio format was altered by presenting probability as either a '1-in-X' ratio (e.g., 1 in 7) or a 'N-in-X*N' ratio (e.g., 10 in 70). Two scenario versions were created for this. In accordance with the literature to date (Pighin et al., 2011; Pighin et al., 2015; Sirota et al., 2014; Sirota et al., 2018a; Sirota & Juanchich, 2019), we expected that participants would overestimate their risk of disease and choose to opt-in to a vaccine more often after observing probability expressed in a '1-in-X' format compared to an 'N-in-X*N' format.

Disease Severity was also manipulated. Two scenario versions were created for this. In one scenario, participants are told about the probability of catching Lyme Disease. This was the low severity condition. In the other scenario, participants are told about the probability of catching Ebola. This was the high severity condition. Given the tendency for the probability of severe outcomes to be inflated (Harris & Corner, 2011), we expected that participants would overestimate their risk of disease and choose to opt-in to a vaccine more often when the outcome was high in severity compared to when the outcome was low in severity.

Finally, we predicted that when participants are estimating likelihood, the effect of Ratio would differ across levels of severity, whereby an overestimation of risk via the '1-in-X' ratio will be greater when used to describe the possibility of a severe outcome. Then, when choosing to accept or refuse a vaccine, there would be an association between ratio and vaccine choice, when severity was both low and high.

5.3.4. Experiment 1 Method

This experiment was pre-registered prior to data collection (see osf.io/xywgm). All materials, raw data and analysis script can be found on the Open Science Framework (OSF, see osf.io/y3kwh/). The aim of our experiment was to test whether two known biases ('1-in-X' bias, 'severity bias') interact with each other in the judgement of medical likelihood. Participants imagined that they were planning a trip abroad, and faced a risk of contracting a disease while on their trip. Risk was presented as either a '1-in-X' ratio (e.g., 1 in 3) or a 'N-in-X*N' ratio (e.g., 13 in 39), and the disease being faced was either low in severity (Lyme Disease) or high in severity (Ebola). As a function of these scenarios, we measured participants' likelihood estimations for catching the disease. We also measured whether participants would accept a safe vaccine against the disease. Our vaccine choice measure was designed in pursuit of extending recent findings, establishing the effects of the '1-in-X' bias on categorical health decisions.

Experiment 1 Design

This experiment had a two-factor (2x2) independent groups design. The first independent variable was Ratio with two levels (1-in-X, N-in-X*N). Participants were randomly assigned a ratio that represented one of four likelihoods, either 7.69% (1 in 13, 7 in 91), 1.20% (1 in 83, 2 in 166), 14.29% (1 in 7, 10 in 70) or 33.33% (1 in 3, 13 in 39). As per our pre-registration, these varying likelihoods were not an experimental factor in this design, rather we averaged scores across likelihoods. The second independent variable was outcome Severity with two levels (High = contracting Ebola, Low = contracting Lyme Disease). Participants were randomly allocated to one of the four experimental conditions (see Table 6).

Table 6

The four experimental conditions in our 2x2 design.

Severity	**1-in-X Ratio**	**N-in-X*N Ratio**
Ebola (High Severity)	1-in-X/High Severity	N-in-X*N/High Severity
Lyme Disease (Low Severity)	1-in-X/Low Severity	N-in-X*N/Low Severity

In total, 16 scenarios were created (4 conditions x 4 likelihoods). We measured participants' estimations of outcome Likelihood, as well as their subsequent Choice to accept or reject the offer of a safe vaccine (Yes vs. No).

Experiment 1 Materials

We created a single experimental vignette for this experiment (see Figure 14), which was manipulated in terms of how the likelihood ratio was presented and how severe the outcome was. The hypothetical scenario asked participants to imagine that they are planning a trip abroad for work or pleasure. They were given information about the chances of contracting a disease while on their trip. Based on this, they were asked to estimate their likelihood of contracting the disease while on their trip, and asked whether they would accept a vaccine against the disease. The likelihood given to each participant was presented as either a '1-in-X' ratio (e.g., 1 in 7) or a 'N-in-X*N' ratio (e.g., 10 in 70). Also, the severity of the disease that could potentially be contracted was either Low (Lyme Disease) or High (Ebola).

Figure 14

Example vignettes.

Low Severity (Lyme Disease)
Lyme Disease is a bacterial infection that can spread to humans by infected ticks. Most people with Lyme Disease get better after antibiotic treatment. In general, the disease is not considered to be severe.

Imagine that you have booked a trip to Sierra Leone in West Africa. You learn from your doctor that the risk of being infected with Lyme Disease during your trip is (1 in 13 / 7 in 91).

High Severity (Ebola)
Ebola virus disease is a serious viral infection that originated in sub-Saharan Africa. Ebola virus disease is often fatal, with half of those infected dying from the disease. In general, the disease is considered to be severe.

Imagine that you have booked a trip to Sierra Leone in West Africa. You learn from your doctor that the risk of being infected with Ebola during your trip is (1 in 13 / 7 in 91).

Ratio Manipulation

Two scenario versions were created to operationalise Ratio within our vignette. In one version, participants observe a likelihood expressed as a '1-in-X' ratio. For example, a 14.29% chance expressed as 1 in 7. In the other version, participants observe a likelihood expressed as a 'N-in-X*N' ratio. For example, a 7.69% chance expressed as 7 in 91 (rather than 1 in 13).

Severity Manipulation

Two scenario versions were created to operationalise Severity within our vignette. In one version, participants are told by their doctor that they have a chance of catching Lyme Disease (Low severity). In the other version, they have a chance of catching Ebola (High severity). In both versions participants are given information on their respective diseases,

which includes a description of the disease and an indication of the consequences of contracting such a disease (see Figure 14).

To ensure the manipulation was perceived as intended, participants were explicitly told the severity of the disease. In the high severity condition (Ebola), the information stated that '*In general, the disease is considered to be severe*'. By contrast, in the low severity condition (Lyme Disease), the information stated that '*In general, the disease is* ***not*** *considered to be severe*'.

Experiment 1 Measures

Likelihood Estimate

After reading the scenario, participants were asked '*In your opinion, what is the probability that you will be infected by Ebola [Lyme Disease] during your trip?*'. Subjective estimations of likelihood were provided as a percentage, on a slider scale ranging from 0 to 100. To ensure that responses provided were in-fact an estimate, participants were explicitly told not to calculate their answer.

Vaccine Choice

To establish vaccine choice, participants were asked '*Considering your chances of being infected by Ebola [Lyme Disease] on your trip, if your doctor offered you a vaccine against the disease that was considered to be safe, would you accept this vaccine?*'. Choice was indicated with 'Yes' or 'No'.

Experiment 1 Procedure

This online experiment was run using Qualtrics survey platform (Snow & Mann, 2013). The experiment was accessed via an anonymous link. After providing basic demographic information, participants saw a scenario preface that reads as the following:

"On the following page you will see some information about a disease, followed by a short imaginary scenario and a single question. Please give your intuitive response to this question rather than trying to calculate the answer."*

**Note - 'single question' refers to our Likelihood Estimate measure, our second DV (vaccine choice) was presented on a subsequent page.*

Once instructed, participants were randomly assigned to one of 16 scenarios (4 possible conditions x 4 possible likelihoods; see osf.io/xgjua/ for an overview of scenarios). After reading their scenario, participants estimated their likelihood of catching the disease. Then, on a new page, they indicated whether they would accept or refuse a vaccine against the disease. Before submitting their response, participants were asked whether they took the survey seriously (Aust et al., 2013), as well as whether they calculated their answer to the likelihood question. The end-of-survey message included debrief information. Median completion time for the entire survey was 1.8 minutes. A .qsf file for the Qualtrics survey can be accessed via the OSF (see osf.io/u7md4/), which can be used to exactly replicate the experiment.

Experiment 1 Participants

Our minimum sample was calculated using the least powerful set of tests (χ^2 tests, df = 1). We conducted a prospective power analysis using the 'pwr' package in R (Champely, 2020). Minimum sample was determined based on the least powerful set of tests. This was our analysis for Vaccine Choice. To detect a medium sized effect with 80% power (alpha = .025) we required a minimum sample of 212 participants.

Participants were recruited via opportunity sampling. The experiment was advertised online through social media and university mailing lists. Undergraduate psychology students could claim course credit for completion. We recruited an initial sample of 905 consenting volunteers. We excluded those who did not complete the experiment (n = 63), those who declared that they calculated an exact answer to the likelihood question (n = 142), and those who declared a non-serious response (n = 3). The final sample comprised 697 English-speaking adults (115 Men, 563 Women, 11 Non-Binary, 3 Prefer not to Say, 5 Self-Identify) aged between 18 and 76 (mean age = 27.2, SD = 10.6). The number of participants allocated to each condition was between 172 and 177.

5.3.5. Experiment 1 Results

The following analysis was pre-registered prior to data collection. The Open Science Framework hosts our pre-registration (see osf.io/xywgm), raw data (see osf.io/rgqam) and analysis script (see osf.io/wgbcv).

All questions required a response, so there was no missing data and no data points were excluded. Analyses were carried out using R version 4.1.0 (R Core Team, 2021) and RStudio version 1.4.1717 (RStudio Team, 2021). An overestimation score for each participant was calculated by subtracting the genuine likelihood from their likelihood estimate. For example, a respondent who saw a 1 chance in 7 (14%), and estimated their likelihood of catching the disease as 18%, provided an overestimation value of 4 percentage points. When analysing both of our measures, we averaged overestimation of likelihood across all four genuine likelihoods (1%, 7%, 14%, 33%). This is firstly in replication of Sirota et al. (2018a), but was also deemed reasonable because a factorial ANOVA (adding genuine likelihood as an IV) revealed that genuine likelihood did not interact with Ratio or Severity.

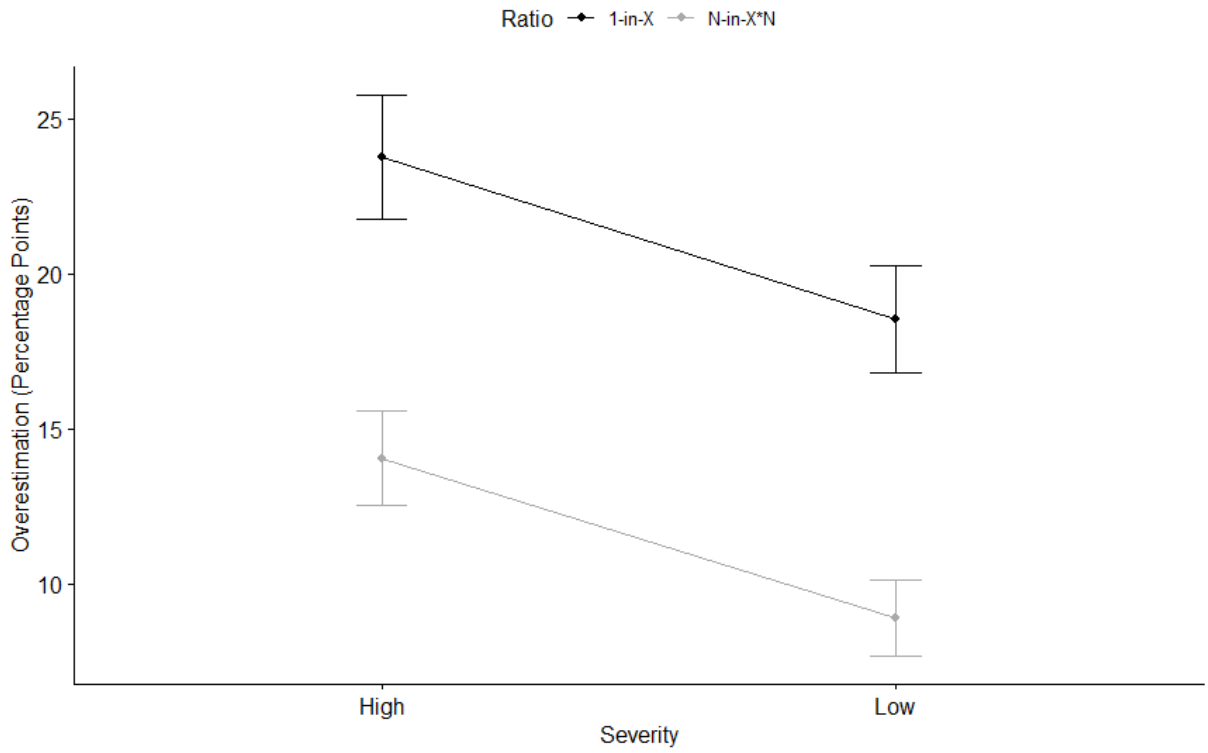
A 2x2 (Ratio x Severity) independent groups ANOVA was conducted to analyse our Likelihood Estimate dependent variable, using the 'aov' base R function. Cohen's *d* effect sizes were calculated for both main effects using the 'psych' package (Revelle, 2021). Our Vaccine Choice dependent variable was analysed with Chi-Square (χ^2) tests of independence, using the 'gmodels' package (Warnes et al., 2018). To determine whether cell counts deviated from their expected counts, Adjusted Standardised (Pearson) Residuals ($\tilde{r}_{ij}$) were calculated alongside each Chi-Square test. Cells that exceed the absolute value = +/-1.96 indicate a significant deviation from the expected count. Phi (φ) is reported as a measure of effect size.

Likelihood Estimation

We measured the extent to which participants overestimated the likelihood of contracting disease, as a function of ratio format and outcome severity. Mean overestimation per condition is displayed in Figure 15.

Figure 15

Line graph displaying mean overestimation (in percentage points) across Ratio and Severity conditions. Error bars represent standard error. These data as a table of Means (SDs) can be accessed via the OSF (see [osf.io/tf5x2/](osf.io/tf5x2/)).



We found a significant main effect of Ratio on likelihood overestimation. Participants who saw the likelihood expressed with a '1-in-X' format (mean = 21.15, SD = 24.30) overestimated their chances of catching the disease significantly more than participants who

saw the likelihood expressed with a 'N-in-X*N' format (mean = 11.44, SD = 18.30), $F(1, 693) = 35.272$, $p < .001$, $d = 0.45$, 95% CI [0.3, 0.6].

We found a significant main effect of Severity on likelihood overestimation. Participants who faced catching the high severity disease (Ebola; mean = 18.9, SD = 23.1) overestimated their chances of catching the disease significantly more than participants who faced catching the low severity disease (Lyme Disease; mean = 13.69, SD = 19.6), $F(1, 693) = 10.111$, $p = .002$, $d = 0.24$, 95% CI [.09, .39]. Finally, there was a non-significant interaction between Ratio and Severity $F(1, 693) = .0003$, $p = .987$. The effect size of the 1-in-X bias was $d = 0.42$ when severity was high and $d = 0.49$ when severity was low.

Vaccine Choice

We hypothesised that there would be an association between Ratio and Vaccine Choice, as well as between Severity and Vaccine Choice. The proportion of yes/no responses across both levels of Ratio and both levels of Severity are presented in Table 7.

Table 7

*Count, Expected Count and Adjusted Standardised Residuals of choice to Accept or Refuse Vaccine, when likelihood is expressed as a '1-in-X' and 'N-in-X*N' Ratio, and when Severity is Low (Lyme Disease) and High (Ebola).*

	Ratio		**Severity**	
	1-in-X	**N-in-X*N**	**High**	**Low**
	Accept Vaccine			
Count	322	326	331	317
Expected Count	320.746	327.254	322.605	325.395
Adjusted Std. Residual	0.372	-0.372	2.487	-2.487
	Refuse Vaccine			
Count	23	26	16	33
Expected Count	24.254	24.746	24.395	24.605
Adjusted Std. Residual	-0.372	0.372	-2.487	2.487

Overall, the large majority of participants chose to accept a vaccine against the disease (92.97%).

Ratio

There was a non-significant association between Ratio and Vaccine Choice $\chi^2(1, N = 697) = 0.138$, $p = .710$, $\varphi = .01$. The proportion of participants who accepted a vaccine did not significantly differ from expected cell counts between the '1-in-X' and 'N-in-X*N' ratio format conditions.

A Chi-Square test analysing the association between Ratio and Vaccine Choice was conducted on each level of Severity (using $\alpha = .025$). There was a non-significant association

between Ratio and Vaccine Choice when Severity was High ($\chi^2(1, N = 350) = .230$, $p = .631$), as well as when Severity was Low ($\chi^2(1, N = 347) = .001$, $p = .972$).

Severity

There was a significant association between Severity and Vaccine Choice $\chi^2(1, N = 697) = 6.188$, $p = .013$, $\varphi = .09$. When the outcome severity was High, 95.39% of participants chose to accept a vaccine against the disease (Ebola), a significantly higher proportion than the expected cell count ($\tilde{r}_{ij} = 2.487$). When the outcome severity was Low, 90.57% of participants chose to accept a vaccine against the disease (Lyme Disease), a significantly lower proportion than the expected cell count ($\tilde{r}_{ij} = -2.487$).

5.3.6. Experiment 2

Experiment 2 of this paper was conducted in response to a surprising finding in experiment 1. Contrary to our hypothesis, we found that ratio format did not significantly influence our participants' decision to accept a safe vaccine against the disease. This does not align with recent experiments which have demonstrated that participants are more likely to make preventative health-based decisions as a consequence of the '1-in-X' ratio, both on a continuous decision measure (e.g., likelihood of cancelling trip abroad; Sirota et al., 2018a) as well as a dichotomous choice task (e.g., yes/no decision to cancel trip abroad; Sirota & Juanchich, 2019). Whilst unexpected, we reasoned as a possible explanation, that the '1-in-X' bias may be reliant upon a balance of risk.

Irrespective of whether participants faced Ebola (high severity) or Lyme Disease (low severity), almost 93% of participants chose to accept the vaccine. Comparatively, in previous research, the proportion of participants deciding to cancel their trip (rather than face the possibility of contracting a disease) was closer to ~40% as reported by Sirota and Juanchich

(2019, p. 34). It may therefore be the case that sensitivity to the '1-in-X' bias is neutralised as the balance of risk is weakened. For example, if a person must decide between (1) risk contracting a disease, or (2) cancelling their trip, they may be more likely to pay attention to their actual risk of catching the disease. By contrast, if a person must decide between (1) risk contracting a disease, or (2) accepting a safe vaccine, they may be less likely to pay attention to their actual risk because the protection of a vaccination dramatically reduces that risk (e.g., a person under 1 in 7 chance of catching the disease is no longer under that risk once they have been vaccinated). In short, ratio format may matter less when the decision is comparatively easy. This interpretation is supported by our significant association between severity and choice. Participants accepted a vaccine more often when the disease was severe, evidencing that the decision is based, at-least in part, on a balance of risk.

To address the above issue, we created a new experimental vignette (adapted from Sirota & Juanchich, 2019) designed to give participants a more difficult health-based decision to make. The aim was to collect follow-up data and test whether the '1-in-X' bias and the severity bias interact to influence dichotomous health-based decisions. Participants imagined themselves as a healthcare patient who has recently developed the condition 'reactive arthritis'. To manage the pain associated with their condition, they are offered a course of steroid medication (prednisolone) that they choose to either accept or refuse. Importantly, the medication has a potential side effect attached to its use.

We again manipulated both ratio format and severity within this scenario. The chance of experiencing the side effect was represented as either 1 in 6 (1-in-X ratio) or 7 in 42 (N-in-X*N ratio), and the potential side effect itself was either high in severity (severe acne) or low in severity (mild acne). Based on our methodological improvements, we predicted that there

would be an association between ratio format (1-in-X, N-in-X*N) and treatment decision (accept, refuse) when side effect severity was both low and high.

5.3.7. Experiment 2 Method

As in experiment 1, this experiment was pre-registered prior to data collection (see osf.io/3b6v8), with all materials, raw data and analysis script freely available on the OSF (see osf.io/smf9k/). Participants imagined that they had been offered a treatment by their doctor, and that the treatment came with the risk of a side effect. We again manipulated both Ratio format and side effect Severity. Risk was presented as either a '1-in-X' ratio or a 'N-in-X*N' ratio (e.g., either 1 in 6 chance or 7 in 42 chance), and the potential side effect was either low in severity (mild acne) or high in severity (severe acne). Based on these conditions, we measured whether participants would accept or refuse the treatment that they had been offered.

Experiment 2 Design

The same two-factor (2x2) independent groups design was utilised. Our first independent variable was Ratio with two levels (1-in-X, N-in-X*N). Our second independent variable was Severity with two levels (High = Severe Acne, Low = Mild Acne). As in experiment 1, participants were randomly allocated to one of our four experimental conditions (see Table 8).

Table 8

Experiment 2 conditions (2x2 design).

Severity	**1-in-X Ratio**	**N-in-X*N Ratio**
Severe Acne (High Severity)	1-in-X/High Severity	N-in-X*N/High Severity
Mild Acne (Low Severity)	1-in-X/Low Severity	N-in-X*N/Low Severity

Our single dependent variable in this experiment was the participants' decision to either accept or refuse their doctor's recommended treatment (Yes vs. No). We also measured how serious participants considered their given side-effect to be. This was done to verify that our manipulation of severity was effective.

Experiment 2 Materials

We created a single experimental vignette for experiment 2 (see Figure 16). The scenario asked participants to imagine themselves as a healthcare patient who has recently developed Reactive Arthritis. To treat this, their doctor recommends a course of steroid treatment (Prednisolone) which will reduce pain but may cause a side effect. Participants are required to make a decision about whether they would accept the recommended treatment. The raw likelihood of getting the side effect was 16.67% and was the same for each participant, presented as either a 1 in 6 chance (1-in-X) or a 7 in 42 chance (N-in-X*N). Unlike experiment 1, we did not average scores across various raw likelihoods (e.g., 1 in 3, 1 in 13, 1 in 82). This is because our results in experiment 1 did not differ based on raw likelihood, suggesting that capturing a range of raw likelihoods is not necessary here. In other words, a single '1-in-X' ratio is sufficiently representative of a range of '1-in-X' ratios. Finally, the severity of the potential side effect was either Low (mild acne) or High (severe acne). Acne is a genuine side effect of the steroid treatment Prednisolone, and, to allow the participants' decision to be a true 'balance of risks', it was important that the challenges posed by the side effect were sufficiently different to the challenges posed by the patient's condition (Reactive Arthritis). For example, without the treatment the participant faces non-visible pain, whereas with the treatment the participant replaces non-visible pain with visible skin problems.

Figure 16

Experiment 2 vignette (adapted from Sirota & Juanchich, 2019).

Reactive arthritis can cause you to have extremely painful, swollen joints and can make you feel very tired. Unlike other types of inflammatory arthritis, for many people reactive arthritis lasts a relatively short amount of time. In most cases, it clears up within a few months.

Imagine that you are a healthcare patient and you have recently developed reactive arthritis. Your symptoms include joint pain, swelling, watery eyes and lower back pain. You cannot take ibuprofen to manage the pain because of a stomach condition, so instead, your doctor suggests that you take a steroid medication called prednisolone.

You ask about potential side effects, and your doctor tells you that there is a [1 in 6/7 in 42] chance of developing [mild/severe] acne while taking prednisolone.

Ratio and Severity Manipulation

The way that we manipulated both Ratio format and Severity are the same as experiment 1. In this experiment however, we did not explicitly tell participants whether the side effect was severe. Instead, we measured participants' perception of how serious they consider their given side effect to be.

Experiment 2 Measures

Treatment Choice

To establish treatment choice, participants were asked '*Considering your chances of getting this side effect, would you accept a course of prednisolone?*'. Choice was indicated with 'Yes' or 'No'.

Manipulation Check (Severity)

To establish perceptions of side effect severity, after reading the scenario participants were asked '*In your opinion, how serious is 'mild [severe]' acne?*'. Subjective ratings of

seriousness were provided on a slider scale ranging from 0 (Not at all Serious) to 10 (Extremely Serious).

Experiment 2 Procedure

As in experiment 1, a Qualtrics survey was accessed via anonymous link. Participants were again briefed with the following preface:

"On the following page you will see some information about a disease. You will then be given a short hypothetical scenario where you are asked to imagine yourself as a healthcare patient. Please read the scenario and answer the 2 questions that follow."

After this brief, participants were randomly assigned to one of four scenarios (2x2 conditions). After reading the scenario, participants indicated their treatment choice and then rated the seriousness of their side effect on the following page. Participants told us whether they took the task seriously (Aust et al., 2013) before being given debriefing information in the end-of-survey message. Median completion time was 1.57 minutes (see osf.io/sfdga for a .qsf file of the survey).

Experiment 2 Participants

We again conducted a prospective power analysis for experiment 2. The criteria for this analysis was identical to experiment 1 and as such, we again required a minimum sample of 212 participants.

Participants were recruited using the same methods as experiment 1. We recruited an initial sample of 654 consenting volunteers. After removing those who did not complete the experiment (n = 29) and those who provided a non-serious response (n = 4), our final sample consisted of 621 English-speaking adults (138 men, 467 women, 8 non-binary, 8 self-identify)

aged between 18 and 77 years (mean age = 29.9, SD = 10.6). The number of participants allocated to each experimental condition ranged from 153 to 158.

5.3.8. Experiment 2 Results

This analysis was pre-registered prior to data collection. The raw data and analysis script for this experiment are freely available on the OSF (see osf.io/fbkpc for data, see osf.io/t4ykh for script). There was no missing data and analyses were carried out in the same software version as experiment 1. Our manipulation check was analysed with a t-test, using the base R function. As in experiment 1, our treatment choice DV was analysed with a series of 2x2 Chi-square (χ^2) tests of independence, using the 'gmodels' package (Warnes et al., 2018). Phi (φ) is reported as a measure of effect size.

Manipulation Check

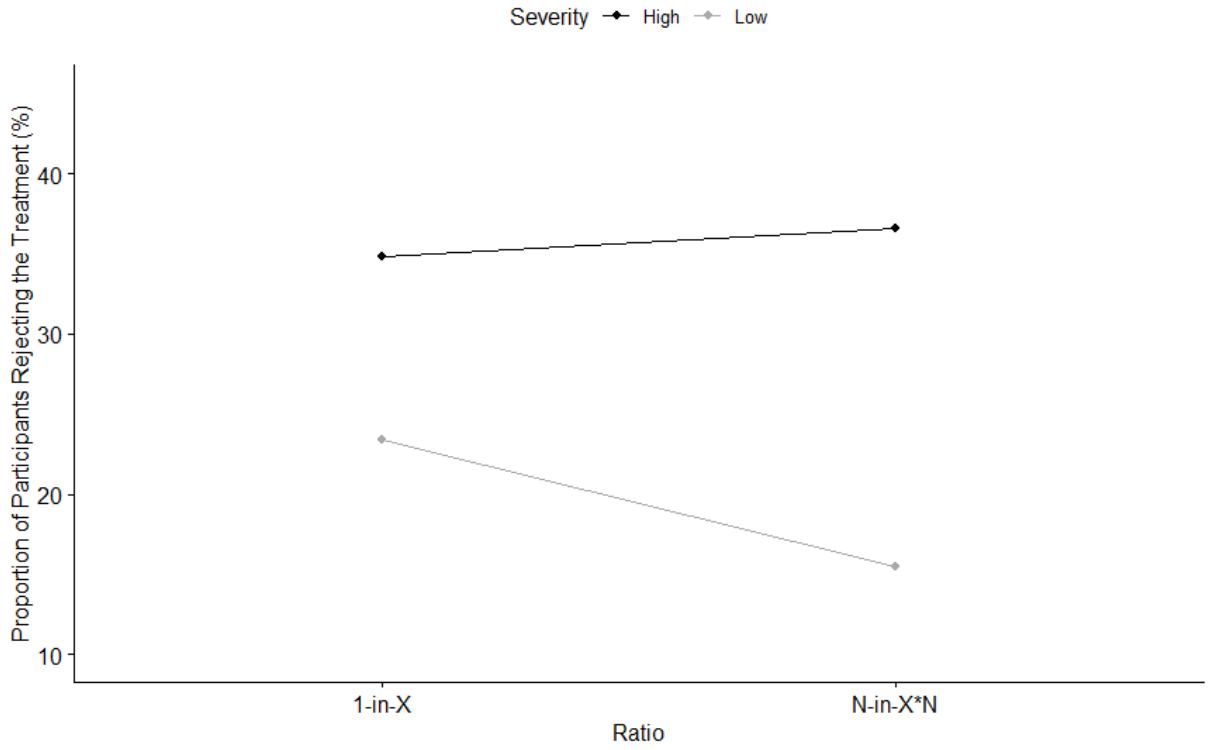
We measured how serious participants considered their given side effect to be. As expected, our participants considered severe acne to be significantly more serious (M = 5.94, SD = 1.96) than mild acne (M = 3.42, SD = 2.01), $t(618.98) = 15.851$, $p < .001$, $d = 1.27$, 95% CI [1.09, 1.46]. This confirmed to us that our manipulation of severity was effective.

Treatment Choice

We again hypothesised that there would be a significant association between Ratio and Treatment Choice, as well as between Severity and Treatment Choice. In Figure 17 we have plotted the proportion of participants who chose to refuse the treatment across our four conditions. That is, the percentage of participants in each condition who chose not to accept the treatment offered by their doctor.

Figure 17

*Line graph showing the proportion (%) of participants who chose to refuse the recommended treatment rather than accept, as a function of ratio format (1-in-X, N-in-X*N) and side effect severity (High, Low).*



Ratio

We found a non-significant association between Ratio and Treatment Choice $\chi^2(1, N = 621) = 0.747$, $p = .387$, $\varphi = .01$. The proportion of participants who refused the treatment did not significantly differ from expected cell counts between the '1-in-X' ratio and 'N-in-X*N' ratio format conditions.

We analysed the association between Ratio and Treatment Choice when Severity was both Low and High individually (using $\alpha = .025$). There was a non-significant association between Ratio and Treatment Choice when Severity was High ($\chi^2(1, N = 308) = .104$, $p = .747$, $\varphi = .01$), as well as when Severity was Low ($\chi^2(1, N = 313) = 3.139$, $p = .076$, $\varphi = .08$).

Severity

There was a significant association between Severity and Treatment Choice $\chi^2(1, N = 621) = 20.482$, $p < .001$, $\varphi = .18$. When side effect Severity was High (severe acne), 35.7% of participants chose to refuse the treatment. This was a significantly higher proportion than the expected cell count ($\tilde{r}_{ij} = 4.526$). When side effect Severity was Low (mild acne), 19.5% of participants chose to refuse the treatment. This was a significantly lower proportion than the expected cell count ($\tilde{r}_{ij} = -4.526$).

5.3.9. Discussion

The aim of this paper was to extend the findings of recent papers (Sirota et al., 2018a; Sirota & Juanchich, 2019). Our first experiment set out to establish whether two known biases ('1-in-X' bias, severity bias) interact with one another to influence estimations of health risk. We successfully replicated both biases, evidencing that both '1-in-X' ratio formats and high severity outcomes induce a greater overestimation of likelihood. However, the biases did not significantly interact, suggesting that they may influence perceptions of health risk independently. On an additional measure of dichotomous health-based decision making, a severe outcome caused our participants to accept a safe vaccine more often, but ratio format (i.e., a 1-in-X ratio) did not influence the decision. Our second experiment re-tested the observed effects on health-based decision making under a different set of circumstances. We observed the same pattern of results as experiment 1, suggesting that 1-in-X ratios may not influence decisions made about health, when they are considered alongside the severity of an outcome.

By replicating recent experiments, our findings from experiment 1 provide additional support for the '1-in-X' effect as a form of cognitive bias. When asked to make an estimate of

probability, people systematically overestimate the likelihood of an outcome expressed by a '1-in-X' format, relative to both objective probability as well as the same probability expressed by a 'N-in-X*N' format (Oudhoff & Timmermans, 2015; Pighin et al., 2011; Sirota et al., 2014; Sirota et al., 2018a; Sirota & Juanchich, 2019).

We also evidence here using a medical scenario, that overestimation of likelihood is larger when the medical outcome is severe (i.e., chance of Ebola vs. chance of Lyme Disease). This is in keeping with a known severity bias whereby more severe health outcomes are judged to be more probable (Bonnefon & Villejoubert, 2006; Fischer & Jungermann, 1996; Weber & Hilton, 1990). As a consequence, our results lend further support to the argument that '1-in-X' ratios should be eliminated from the communication of health risks to patients (Zikmund-Fisher, 2011, 2014), because they demonstrate that raw overestimation of risk is at its highest under more severe circumstances. In other words, the perceived risk that a '1-in-X' ratio represents, is at its most inaccurate when the consequences for the patient are at their greatest (e.g., breaking bad news, Baile et al., 2000; Berkey et al., 2018).

In Experiment 1, participants in the '1-in-X' condition considered the likelihood of catching a disease to be 21.15 percentage points above the genuine probability. To compound this, participants who faced the high severity disease estimated their likelihood of catching this disease to be 5.21 percentage points above those who faced the low severity disease. The use of '1-in-X' ratios when communicating health risks to patients may therefore result in an inflated perception of medical likelihood, particularly when the outcome is worrisome or extreme in nature. This has important implications considering that '1-in-X' ratios are largely favoured as a communicative strategy by health professionals, particularly when the information exchange is negative (Sirota et al., 2018b).

Interestingly though, our non-significant interaction suggests that the degree of overestimation is not especially influenced by a particular combination of ratio format and severity level. That is, when severity was low and high, there was a medium sized effect of ratio on overestimation (d = .49 and d = .45 respectively). Put another way, the difference in overestimation between the '1-in-X' and 'N-in-X*N' ratio format was not made worse by a particularly severe outcome. This was an unexpected finding because, as theorised by Sirota et al. (2018a), it was thought that the same cognitive mechanism that activates to cause the 1-in-X bias is the same mechanism that activates when judging the likelihood of a severe outcome. For this reason, it was hypothesised that a severe outcome would effectively exaggerate (or worsen) the 1-in-X bias. Instead, our data suggests that an overestimation of risk is caused when a '1-in-X' ratio is used, and when the outcome is severe, but whilst the two processes occur simultaneously, they do so independently. Despite not matching our expectations, there may be positive practical implications attached to this finding, because it could simplify the process of phasing out the '1-in-X' format within health communication. Specifically, if the use of '1-in-X' ratios were to be eliminated then the accuracy of likelihood communication would improve across-the-board, rather than in only particular circumstances (e.g., when breaking bad news).

In Experiment 1, exposure to a '1-in-X' ratio did not influence the decision to accept a safe vaccine against a potential disease. For example, the same number of participants accepted a vaccine against a 1 in 7 chance of catching Lyme Disease, as they did against a 10 in 70 chance of catching Lyme Disease. In experiment 2 we observed this effect when the decision involved notably negative outcomes on both sides (i.e., a painful symptom vs. the chance of a negative side-effect). In keeping with experiment 1, exposure to a '1-in-X' ratio in experiment 2 did not significantly influence the decision to accept or refuse a recommended

treatment. Therefore, we must consider the possibility that the 1-in-X bias does not influence particular health-based decisions, a finding that is in contrast with recent work (Sirota et al., 2018a; Sirota & Juanchich, 2019). Below we put forward a possible explanation for this result.

Perhaps it is the case in experiment 2, that the prospect of a particularly severe side effect is 'overpowering' the relevance of a '1-in-X' ratio, when making a decision based on these circumstances. When faced with *severe acne* (high severity), there was no meaningful difference in the proportion of participants refusing the treatment between the '1-in-X' and 'N-in-X*N' conditions (34.8% vs. 36.6%). By contrast, there was a larger raw difference in the proportion of participants refusing the treatment, when faced with *mild acne* (low severity) between the '1-in-X' and 'N-in-X*N' conditions (23.4% vs. 15.5%). We must be cautious here because the difference under high severity was non-significant ($p = .076$), however, the raw pattern of data could point to a prioritisation of outcome severity over ratio format. In other words, when faced with a particularly severe outcome, participants placed more 'weight' on this element of the utterance (rather than the ratio format) in determining whether they should accept or refuse the recommended treatment. Effectively, the severity of the outcome is considered a more immediate threat than the represented likelihood. This could suggest that when considered in isolation, '1-in-X' ratios have a significant impact on health-related decisions, but when tested under threatening circumstances, the relevance of this ratio format is weakened. Establishing the extent to which this is true would be a worthwhile pursuit for future research.

Finally, across both studies we discovered that a more severe outcome had a significant impact on decision-making. In experiment 1, participants accepted a safe vaccine more often to protect against a higher severity disease (Ebola vs. Lyme Disease). In experiment 2, more participants refused a treatment when the potential side effect was higher in severity (severe

vs. mild acne). These findings simply reflect that people make health-based decisions in the interest of avoiding serious negative outcomes.

Evaluation and Conclusions

Our experiment is valuable for a number of reasons. Firstly, our findings replicate and extend previous research by evidencing that both ratio format and severity influence the overestimation of health risk on an individual basis. Consider again a statement that posits a '1-in-X chance of Y'. We demonstrate that the use of a '1-in-X' ratio (rather than a 'N-in-X*N' ratio) causes an overestimation of the likelihood of Y. Moreover, we simultaneously demonstrate that the outcome Y is considered to be more likely when severe.

Secondly, our measures of decision-making are insightful as we focus on the decision to opt in or out of a specific treatment or intervention. Arguably, measuring in this way is more directly applicable to doctor-patient communication, as health professionals are more likely to communicate risks associated with medications or treatments (Fagerlin, Zikmund-Fisher & Ubel, 2011).

On the other hand, the most obvious factor limiting our experiments here is the fact that our vignette is hypothetical, so we are unable to extrapolate these findings to *literal* interpretations of actual speech. This is less of an issue as it pertains to the perceptions of probability, because a number of studies have evidenced that there are no large differences between actual and hypothetical situations in this regard (Pighin et al., 2013; Pighin et al., 2015). However, there is less direct evidence for the study of health-based decisions, so the effects tested here would be better validated using a measure of genuine (rather than hypothetical) health-based decisions.

Overall, this paper was able to successfully replicate the effects of both the '1-in-X' bias and the severity bias on perceptions of medical likelihood, whereby the use of a '1-in-X' ratio (e.g., 1 in 7), and a severe outcome (e.g., catching Ebola) caused inflated overestimations of how likely the event was perceived to be. The biases did not significantly interact, suggesting that they both influence likelihood perception individually, and that if they were eliminated from health communication, the improvements in communicative accuracy would not be limited to only particular circumstances. As expected, a high severity outcome caused a larger proportion of participants to take evasive action (e.g., refuse treatment). Unexpectedly though, ratio format across both studies did not influence our participants' decisions. Differences in our approach compared to other studies could point towards a 'dampened' sensitivity to the '1-in-X' bias when a decision-maker is faced with a severe medical outcome. Our complete findings support calls to retire the '1-in-X' format as a method of communicating health risks to patients.

Acknowledgements: We thank Isabel Dunkerley who worked with us as a research assistant to assemble Experiment 2 prior to data collection.

Author Contribution Statement

Study Conception: Harry Clelland, Matthew Haigh

Study Design: Harry Clelland, Matthew Haigh

Development of Materials: Harry Clelland, Isabel Dunkerley

Pre-Registration: Harry Clelland, Isabel Dunkerley

Data Collection: Harry Clelland, Isabel Dunkerley

Data Analysis: Harry Clelland

First Draft: Harry Clelland

Review and Editing: Harry Clelland, Matthew Haigh

5.4. Chapter Summary

Across four experiments reported in this chapter, we set out to answer our second research question: '*How do pragmatic factors influence the production and interpretation of quantifiers?*'. Our first experiment tested whether face-threat could influence the production of a proportional quantifier like 'some' or 'many'. Our second experiment tested whether a severe outcome could influence the interpretation of these terms. Finally, our third and fourth experiment tested whether a severe outcome would mediate the overestimation of risk caused by '1-in-X' ratios. The paragraph below outlines how the findings reported in this chapter directly answer our second research question (RQ2).

We discovered that high face-threat significantly influences the production of quantifiers, and causes speakers to be more motivated by face-management (rather than accuracy). These findings advance the current understanding of politeness and face-work, by demonstrating that speakers can alter communicated probability under high face-threat with words such as 'some' and 'many', in the same way that probabilistic phrases can be altered to manage face (e.g., 'somewhat unlikely'; Sirota & Juanchich, 2015). A high severity outcome causes hearers to consider the use of words like 'some' or 'many' to be a politeness strategy, but these terms do not make a high severity outcome seem more likely to occur. Lastly, a high severity outcome causes an overestimation of risk in the same way that a '1-in-X' ratio does, however, severity does not mediate the '1-in-X' bias. Opposing previous research, '1-in-X' ratios did not cause a change in medical decision-making. This suggests that the influence of these ratios in the decision-making process is more multifaceted than was once thought.

In chapter six which follows, we test the pragmatic factors that influence generic language. Unlike uncertainty terms and quantifiers, which are typically used when avoiding

harm to negative face (i.e., used to avoid damaging the self-esteem of either the speaker or the hearer), generic language is useful for readers and writers looking to (intentionally or unintentionally) bolster their positive face. In other words, generics tend to 'prop up' a person's self-image because they make one's work appear more valuable (DeJesus et al., 2019). For this reason, generics are more directly applicable to health communication within the media, primarily because generic statements (e.g., '*therapy helps young people*') are used to refer to population-level tendencies rather than individual circumstances. Chapter six explores how contextual differences impact the production of generic generalisations, as well as the interpretation of these statements compared to their non-generic counterparts (RQ3).

Chapter 6. How do Pragmatic Factors Influence the Production and Comprehension of Generic Language?

6.1. Generic Health Research Claims are Associated with the Generalisability of the Data, but Also the Writer's Motivation to be Appealing (Exp. 7)

6.1.1. Abstract

Generic statements are frequently used to summarise academic research findings (e.g., "Music helps Alzheimer's patients"). A consequence of generic language is that it can imply universal rules about a category or group (e.g., "Music helps *all* Alzheimer's patients") even when the claim is only true for some members of that group. This experiment sought to establish whether a writer's motivation, as well as the generalisability of data being used to make a claim, could influence their decision to choose generic language over a qualified alternative. Participants were asked to caption a medical information article by choosing either a generic claim ("New supplement boosts immune system function in men") or a qualified claim ("New supplement boosts immune system function in *some* men"). They were given data on the effectiveness (generalisability) of the intervention (either 25%, 50%, 75% of sampled men benefitted from the supplement). We also manipulated motivation, with participants motivated to choose either the 'most accurate' or the 'most appealing' caption. As predicted, more participants chose the generic caption when motivated to be appealing, and the proportion choosing the generic caption increased as generalisability increased. When

motivated to be accurate, the majority of participants used generic language only when the data had a high degree of generalisability (75%), but when motivated to be appealing the majority used generic language even at the lowest level of generalisability (25%). Generic language may not imply a high degree of generalisability when the writer is motivated to be appealing. This has implications for the primary and secondary reporting of academic research, where there is ever increasing pressure to be appealing.

Keywords: Generics, Scalar Implicature, Communication, Experimental Pragmatics, Media, Scientific Consensus

6.1.2. Introduction

A generic statement is a piece of language that expresses a generalisation about a category (Cimpian, Brandone & Gelman, 2010). For example, consider statement (1):

(1) Men father children

The generic utterance (1) expresses a characteristic (fathering children) on the basis of a group identity (men). An important property of generic statements is that they have a distinct tolerance to exceptions (Krifka et al., 1995). In other words, they are difficult to prove false, meaning that a person would rarely contest the truthfulness of a generic proposition. By way of demonstration, consider the generic statement (1) compared to its universal equivalent (2):

(2) *All* men father children

Unlike the universal statement (2), the generic statement (1) can still be uttered truthfully even in the face of obvious exceptions, such as the common knowledge that some men cannot father children, and some men choose not to become a father. However, the universal statement (2) cannot be true when there are exceptions. A single man who does not father any children is evidence enough to conclude that '*not all men father children*', falsifying the universal (2) (Lazaridou-Chatzigoga, 2019).

The 'resistance-to-exceptions' that generics possess make them an inviting linguistic option for writers, because they allow ostensibly universal claims to be made without the need for wholly universal evidence. In a series of four experiments, Cimpian et al. (2010) compared interpretations of generic statements (e.g., "Lorches have purple feathers") to a quantified version of the same statement (e.g., "Most lorches have purple feathers"). Their results showed that when asked to estimate characteristic prevalence, the generic statement implied to participants that the properties occur in over 90% of category member cases (i.e., over 90% of

lorches have purple feathers), significantly more than the quantified statement. Moreover, another experiment found that even when participants had the knowledge of a low prevalence (i.e., knowing that less than 50% of lorches have purple feathers), the generic statement was often still judged as true. Taken together, these findings suggest that generic language can be used to make broad generalisations about a category (e.g., 'Lorches', 'Lions', 'Men', 'Women', 'Children' etc.), even when the claim is not true for all members of the category. Generics may therefore be favoured over non-generics, because they allow writers to make striking and easily digestible claims (e.g., 'men father children') that are robust against counterexamples (e.g., 'many men do not father children').

Generic language is common in everyday social discourse and recently it has been demonstrated that is also common in scientific discourse. DeJesus et al. (2019) measured how often generic statements were used by authors among a sample of >1000 psychological research papers. They coded the use of generic language in the three most-read sections of journal articles (Title, Highlights, Abstract; Pain, 2016). The authors found that almost 90% of codable titles included a generic statement (e.g., '*group discussion improves lie detection*'), a much larger proportion than highlights and abstracts. These findings suggest a clear preference for generics when summarising research findings, particularly in more word-limited formats (i.e., titles).

The popularity of generic language in this domain may occur for at least two reasons. The first is because authors are generally *motivated* to make their work appealing (to make clear, bold, memorable claims about the significance and impact of their work), sometimes at the expense of being entirely accurate (e.g., "Men benefit from testosterone therapy" rather than "Some men who took part in a single trial benefitted from testosterone therapy"). The second, is because generic language allows authors to efficiently *generalise* the findings from

their specific sample to the wider population. For example, rather than stating "*x improved y among some of the men in our sample of volunteers*" the aim of most inferential research is to make more general claims about the population, which can be efficiently expressed using generic language (e.g., "*x improves y in men*"). Here follows a more detailed consideration of the influences of motivation and generalisability for producing generic language.

First, the use of generics may aid in the relative success of an article, potentially making writers motivated to utilise them. Dumas-Mallet et al. (2018) reports that journal articles with 'catchy' titles are picked up more often and are treated more favourably by the news media, whereas articles that address the limitations of a sample (i.e., replication statements) were far less 'hyped'. Moreover, progressively shorter distribution formats (e.g., Twitter, blogs) increase pressure on writers to make bold and 'snappy' claims about the implications of their research (Weinstein & Sumeracki, 2017). Generic titles provide the means to imply widely applicable conclusions in a format that satisfies the demand for concise and easy-to-understand research. Those both writing primary (i.e., academics) and secondary (i.e., journalists) research reports may therefore be motivated to favour generic claims above more accurate alternatives.

Second, generics amplify the apparent generalisability of research findings. Growing evidence suggests that generics can be used to consciously or unconsciously make findings appear more generalisable (universally applicable) than what can truly be inferred from the data (Cimpian et al., 2010; DeJesus et al., 2019; Gelman & Roberts, 2017; Gutiérrez & Rogoff, 2003; Leslie, 2007; Leslie 2008; Rogoff, 2003). For example, a male may read a research summary that states 'Caffeine improves barrier function in male skin' (Brandner et al., 2006) and presume that caffeine improves skin barrier function in *all* males. In this case, the reader believes that the statement is highly generalisable, because stating that 'caffeine

improves barrier function in male skin' is not disproven by a single case of non-improved male skin. As such, generics may allow a writer to imply their findings are universal, even in the absence of universal evidence. Improved barrier function would need to be observed in 100% of men in order to truthfully state that 'caffeine improves barrier function in *all* male skin', whereas comparatively, a smaller proportion of men seeing improved barrier function is required to state that 'caffeine improves barrier function in male skin'.

Whilst studies of generic comprehension are common in the literature, experimental studies of generic language production are relatively rare. Recently, Haigh, Birch and Clelland (2022) tested the likelihood of producing a generic statement based on the respective generalisability of scientific opinion (i.e., expert consensus). In their experiment, participants read a news article about a whale washing up onshore. They were told that an independent survey of marine experts was conducted to collect opinion on what caused the whale to die. Expert consensus was reported as a percentage, with participants assigned to one of 20 consensus conditions ranging from 5% agreement to 100% agreement (e.g., "An independent survey found that [70%] of marine experts believe that the whale's death was caused by hitting a commercial ship"). After reading the scenario, the task was to select the most appropriate headline to caption the story from one of two options; either a generic (e.g., Experts believe North Essex whale died after impact with ship) or a qualified non-generic (e.g., Some experts believe North Essex whale died after impact with ship). Results showed that as expert consensus increased, the proportion of participants choosing the generic headline increased. The majority of participants chose the generic headline when consensus reached 70%. These findings suggest that willingness to produce generic language increases as a function of data generalisability (i.e., objective generalisability is positively associated with the likelihood of

using generic language), however, a relatively high degree of consensus (~70%) is required before the majority of individuals choose to use a generic statement.

This generic production experiment by Haigh et al. (2022) demonstrates how the generalisability of data can influence a writer's decision to use generic language. This, however, is unlikely to be the only factor that influences the production of generic claims. As discussed above, the decision to produce a generic statement may also be influenced by the internal motivations of the writer. In this production experiment we focus on the reporting of health related interventions commonly seen in the press (e.g., "Plasma Therapy Helps Critically Ill Covid Patients"; WebMD, 2020). We provided participants with details of a research study and asked them to select the most appropriate headline to caption the summary (either a generic or a qualified headline). Within the summary, we manipulated the objective generalisability of the results (the results were generalisable to either 25%, 50% or 75% of the population) and the motivation of the language producer (to be accurate or to be appealing) to find out how these factors combine to influence the likelihood of using generic language.

6.1.3. The Present Research

Haigh et al. (2022) evidenced that the likelihood of producing generic language increases as the strength of scientific consensus increases. The aim of the current study is to extend these findings, by understanding if the decision to use generic language when summarising health information can increase as a function of a writer's motivation to be appealing, as well as the efficacy (generalisability) of the data being used to state the claim.

We presented participants with a hypothetical scenario. In this scenario, participants were asked to imagine that they were working for a large pharmaceutical company. The company is releasing a medical information article about their new supplement that has been

designed to boost immune system function in men. Participants were given data on the efficacy of the supplement, as well as a brief on the goals of the article. They were then tasked with choosing one of two headlines. They could either select the generic headline ('*New supplement boosts immune system function in men*') or the non-generic headline qualified with 'some' ('*New supplement boosts immune system function in some men*').

We made two manipulations. We manipulated participants' motivation by altering their brief. In one condition, participants were told 'your brief is to select the headline that is most accurate'. In the other condition, participants were told 'your brief is to select the headline that is most appealing'. We expected that a greater proportion of participants would choose the generic headline when motivated to be appealing than when motivated to be accurate. Among psychological literature, generic use is rife within an environment that is increasingly conducive to appealing claims (DeJesus et al., 2019; Dumas-Mallet et al., 2018; Weinstein & Sumeracki, 2017), suggesting that an explicit motivation to be appealing would induce a greater preference for the generic headline in our study.

We also manipulated the generalisability of the data given on the new supplement. There are three levels to this manipulation. Participants were told that their article reports data from a large-scale trial, which shows that either 25%, 50% or 75% of men saw improved health after taking the new supplement. We expected that, in line with the findings of Haigh et al. (2022), the proportion of participants choosing the generic headline would increase as objective generalisability increased. However, we expected that we would only observe this association when participants were motivated to choose the 'most accurate' headline. In other words, we predicted that generic use would increase as generalisability increases, but only when accuracy was the primary motivation. Generics allow a writer to imply universal and timeless conclusions when summarising research data (DeJesus et al., 2019), and is therefore

more appealing than our qualified statement. We expected therefore that when specifically motivated by appeal, the catchier generic headline would be chosen most often irrespective of the level of generalisability.

It was important for our scenario to apply to a health context. Firstly, because studies on human health are one of the most common research topics reported by the media (Bossema et al., 2019; Bratton et al., 2020), meaning that our findings would be largely applicable to the average person. Secondly, the accuracy of public understanding with regards to health research is important. When reading generic statements are attached to 'doctors', 'scientists' or 'experts' some individuals will attribute these statements to *all* doctors, scientists or experts (Haigh et al., 2020). Plus, without the presence of an obvious counterargument, the public will default to assuming a high degree of scientific consensus (Aklin & Urpelainen, 2014). This is despite the fact that across-the-board consensus is rare for clinicians. Consequently, without a proper understanding of generics within a health context, the public remains at risk of encountering misrepresented research summaries, leading to potentially ill-informed health decisions in the absence of alternative information.

Our scenario is based on secondary reporting (i.e., media) in which the results are reported to the general public, rather than primary research reporting (i.e., journal articles), which is written with professional readers in mind. This decision was taken for multiple reasons. Firstly, a lay person is more likely to read news headlines that present short secondary summaries of research. Secondly, this context is more susceptible to manipulations of motivation. If our scenario was to report primary research findings, this would mean the participant acting as an academic, who is (more so than a journalist) obligated to report out of accuracy rather than appeal. In secondary reporting however, it is more conceivable to have a writer motivated more by accuracy, or to have a writer motivated more by appeal. Finally, and

perhaps most importantly, the implications of generics are arguably more consequential when reported in the media. The typical journal article readership consists of academics trained in research methods, meaning they have an understanding of the limitations of a sample (i.e., they understand that a generic summary does not necessarily apply to all cases). Media articles by contrast, are most often read by members of the public who may not have such training and are therefore potentially more susceptible to the inaccuracies of generics.

Collectively, the danger posed by health-related generics in the media is that health-based decisions are taken as a result of exaggerated claims. For example, a parent reading a headline that states 'caffeine helps children' could infer that 'caffeine helps *all* children' and subsequently increase their child's caffeine consumption (Haigh et al., 2020; Leslie et al., 2011). It is therefore crucial to understand the psychological mechanisms that inform a writers' decision to favour a generic over a non-generic statement.

6.1.4. Method

This study was pre-registered prior to data collection (osf.io/3xnrc). All materials, raw data and analysis script can be found on the Open Science Framework (OSF, see osf.io/2wrfq/).

Design

This experiment used a two factor (2x3) independent groups design. The first independent variable was Motivation with two levels (to be Accurate, to be Appealing). The second independent variable was Generalisability of the intervention with three levels of efficacy (25%, 50%, 75% of men). Participants were randomly allocated to one of six possible conditions. The order in which the generalisability and motivation information was presented within the vignette was counterbalanced, so there were two versions of each condition (see

Figure 1). We measured the Headline Choice made by each participant (Generic vs. Qualified Non-Generic).

Participants

We conducted a prospective power analysis using the 'pwr' package in R (Champley, 2020). For our 2x3 chi square omnibus test we required a minimum of 108 participants to detect a medium sized effect with 80% power (assuming alpha = 0.05). However, we also pre-registered three 2x2 planned comparisons. Each of these three comparisons required 117 participants to detect a medium sized effect with 80% power, therefore we aimed to recruit a minimum sample of 351 participants (117x3).

Participants were recruited via opportunity sampling. The study was advertised online through research participation websites (e.g., Prolific, Reddit, Survey Circle), social media and university mailing lists. Undergraduate psychology students could claim course credit for completion and participants recruited via Prolific (n = 42) were compensated £0.20. We recruited an initial sample of 1256 consenting volunteers. We excluded those who did not complete the study (n = 73) and those who declared a non-serious response (n = 6). Our final sample consisted of 1177 English-speaking adults (261 male, 908 female, 8 other) aged between 18 and 85 (Mean age = 29.20, SD = 11.35). The number of participants allocated to each condition was between 194 and 199.

Materials

A single experimental vignette was written for this study (see Figure 18). The hypothetical scenario asked participants to imagine that they worked in the communications team of a large pharmaceutical company. Within this context, they were asked to select the most appropriate headline to caption a medical information article about a new product.

Participants were briefed to choose the headline that was either most accurate or most appealing. They were also given some data on the percentage of the population ("men" in this case) for which the supplement is effective (either 25%, 50% or 75%). Participants were tasked with choosing one of two headlines.

Figure 18

Example vignettes. The order in which the generalisability and motivation information was presented was counterbalanced to cancel out any potential order effects. The presentation order of the two headline choice options was also randomised.

Counterbalance order A (Generalisability information presented before Motivation)

Imagine that you are part of a communication team for a major pharmaceutical company. You have written a medical information article about a new supplement designed to boost immune system function in men. The article reports data from a large-scale peer reviewed trial in which **(25%/50%/75%) of men*** who took part had improved immune system function after 28 days of taking the supplement.

You now need to select an appropriate headline to caption the article. Your brief is to **select the headline that is most (accurate/appealing).**

Which of the following headlines would you choose?

"New supplement boosts immune system function in men"
"New supplement boosts immune system function in some men"

Counterbalance order B (Motivation information presented before Generalisability)

Imagine that you are part of a communication team for a major pharmaceutical company. You have written a medical information article about a new supplement designed to boost immune system function in men. You now need to select an appropriate headline to caption the article. Your brief is to **select the headline that is most (accurate/appealing).**

The article reports data from a large-scale peer reviewed trial in which **(25%/50%/75%) of men** who took part had improved immune system function after 28 days of taking the supplement.

Which of the following headlines would you choose?

"New supplement boosts immune system function in men"
"New supplement boosts immune system function in some men"

*Note - text in bold was also presented in bold to participants.

Motivation Manipulation

To operationalise motivation within our vignette, two versions were created. In one version, participants are briefed '*to select the headline that is most accurate*'. In the other version, participants are briefed '*to select the headline that is most appealing*'. This information was presented in bold text.

Generalisability Manipulation

Participants were informed about data suggesting that the supplement is effective in either 25%, 50% or or 75% of men. This information was presented in bold text.

Measures

Headline Choice

The task was to select an appropriate headline for the article. After reading the generalisability and motivation information, participants were asked "Which of the following headlines would you choose?". The response was dichotomous, as participants could choose one of the following two options:

1. "New supplement boosts immune system function in men"
2. "New supplement boosts immune system function in some men"

Option 1 is a Generic statement. Option 2 is identical, except it is a Non-Generic Qualified with the word '*some*'. Each participant chose one headline. Presentation order of the two choices was randomised to control for order effects.

Open-Ended Headline (Exploratory)

After making their headline choice, participants were given the opportunity to produce their own headline. This was an optional response. Participants read "*If you think there is a better way to phrase the headline, you can write it here in your own words… (optional)*". A text entry box was provided with no minimum or maximum character requirements.

Procedure

This online study was run using Qualtrics survey platform (Snow & Mann, 2013). Participants accessed the study via an anonymous link. After providing basic demographic information, participants saw a scenario preface that reads as the following:

"*In this short survey you will be asked to imagine that you work for a large pharmaceutical company, and that you are making a medical information article about a new supplement.*

Please do your best to immerse yourself in the scenario. Read the scenario once or twice and give the first answer that comes to mind."

On the following page, participants were randomly assigned to one of 12 scenarios (2x3 motivation*generalisability conditions, in two counterbalanced orders 6 motivation-first, 6 generalisability-first; see osf.io/sb73k/ for an overview of conditions). They read the scenario, and chose one of two headlines. Participants could then choose whether or not to write their own alternative headline in a free-text box. Before completion, participants were asked whether they took the survey seriously (Aust et al., 2013). Respondents were informed that they would still be rewarded for their contribution regardless of their answer (e.g., Prolific participants were still paid). The end-of-survey message included debrief information. Median completion time for the entire survey was 1.72 minutes. A .qsf file for the Qualtrics survey can be accessed via the OSF (see osf.io/jk8uz/).

6.1.5. Results

The following analysis was pre-registered prior to data collection. The pre-registration, as well as raw data and analysis script can be found on the Open Science Framework (see osf.io/3xnrc for pre-registration, osf.io/8tmcu/ for data and osf.io/7phxu/ for script).

Questions that addressed the pre-registered analysis required a response, so there was no missing data and no data points were excluded. Analyses were conducted using R version 4.0.2 (R Core Team, 2020) and RStudio version 1.3.959 (RStudio Team, 2020). Both dependent variables were analysed with Chi-Square (χ^2) tests of independence using the 'gmodels' package in R (Warnes et al., 2018). Adjusted Standardised (Pearson) Residuals ($\tilde{r}_{ij}$) were calculated alongside each Chi-Square test to determine whether cell counts deviated from their respective expected counts. Cells that exceed the absolute value = +/-1.96 indicate a significant deviation from the expected count. Phi (φ) and Cramer's V (V) are reported as measures of effect size.

Alternative captions provided in the Open-ended text boxes were categorised by the first author and refined after discussion with the second author. The data with category tallies and content descriptors can be accessed via the OSF (see osf.io/k7u5v/).

Motivation and Headline Choice

We predicted that there would be an association between Motivation (to be 'most accurate' or to be 'most appealing') and Headline Choice (either the Generic statement, or the Qualified statement using 'some'). To test this, we created three subsets of the data dividing the data by Generalisability (25%, 50%, 75%). For example, the first subset included only those in our sample who were given the new supplement scenario with 25% generalisability. A Chi-Square test was conducted on each level of Generalisability (using α = .0167). We

hypothesised that there would be an association between Motivation and Headline Choice at all three levels of Generalisability (Hypothesis 1). Findings are summarised in Table 9, and are visualised in Figure 19.

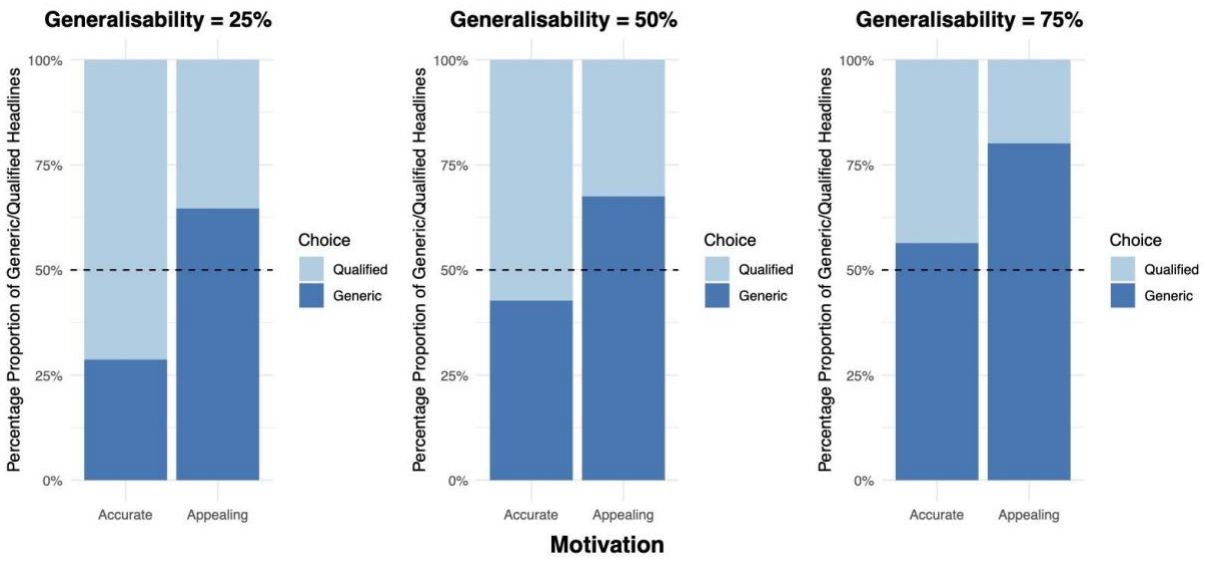
Table 9

Count, Expected Count and Adjusted Standardised Residuals of Generic and Qualified Headline choices, when motivated by Accuracy or Appeal across the levels of Generalisability (25%, 50%, 75%).

	25%		**50%**		**75%**	
	Motivation		**Motivation**		**Motivation**	
	Accurate	**Appealing**	**Accurate**	**Appealing**	**Accurate**	**Appealing**
	Generic Choice					
Count	56	128	85	131	110	157
Expected Count	91.30	92.70	109.37	106.63	113.56	113.84
Adjusted Std. Residual	-7.137	7.137	-4.943	4.943	-5.003	5.003
	Qualified Choice					
Count	139	70	114	63	85	39
Expected Count	103.70	105.30	89.63	87.37	64.84	62.16
Adjusted Std. Residual	7.137	-7.137	4.943	-4.943	5.003	-5.003

Figure 19

Stacked bar charts showing percentage proportion of headline choice (Generic, Qualified) in the accurate and appealing motivation conditions, across the levels of Generalisability (25%, 50%, 75%).



25% Generalisability

There was a significant association between Motivation and Headline Choice $\chi^2(1, N = 393) = 50.934$, $p < .001$, $\varphi = .36$. When motivated to select the 'most accurate' headline, the minority of participants chose the Generic headline (28.72%), significantly less frequently than the expected cell count ($\tilde{r}_{ij} = -7.14$). When motivated to select the 'most appealing' headline, the majority of participants chose the Generic headline (64.65%), significantly more frequently than the expected cell count ($\tilde{r}_{ij} = 7.14$).

50% Generalisability

There was a significant association between Motivation and Headline Choice $\chi^2(1, N = 393) = 24.432$, $p < .001$, $\varphi = .24$. When motivated to select the 'most accurate' headline, the minority of participants chose the Generic headline (42.71%), significantly less frequently than the expected cell count ($\tilde{r}_{ij} = -4.94$). When motivated to select the 'most appealing'

headline, the majority of participants chose the Generic headline (67.53%), significantly more frequently than the expected cell count ($\tilde{r}_{ij}$ = 4.94).

75% Generalisability

There was a significant association between Motivation and Headline Choice χ^2(1, N = 391) = 25.336, p < .001, φ = .25. When motivated to select the 'most accurate' headline, the majority of participants chose the Generic headline (56.41%), significantly less frequently than the expected cell count ($\tilde{r}_{ij}$ = -5.03). When motivated to select the 'most appealing' headline, the majority of participants also chose the Generic headline (80.10%), significantly more frequently than the expected cell count ($\tilde{r}_{ij}$ = 5.03).

Generalisability and Headline Choice

We predicted that there would be an association between Generalisability (efficacy test results indicating 25%, 50% or 75% generalisability) and Headline Choice (either the Generic statement, or the Qualified statement using 'some'), but *only* when participants were motivated to be 'most accurate'.

To test this, we created two subsets of the data. One subset included only those who were motivated to select the 'most accurate' headline. The second subset included only those who were motivated to select the 'most appealing' headline. A Chi-Square test was conducted on both levels of Motivation (using α = .025). We hypothesised that there would be an association between Motivation and Headline Choice, but *only* when participants were motivated to be 'accurate' (Hypothesis 2). Findings are summarised in Table 10, and visualised in Figure 20.

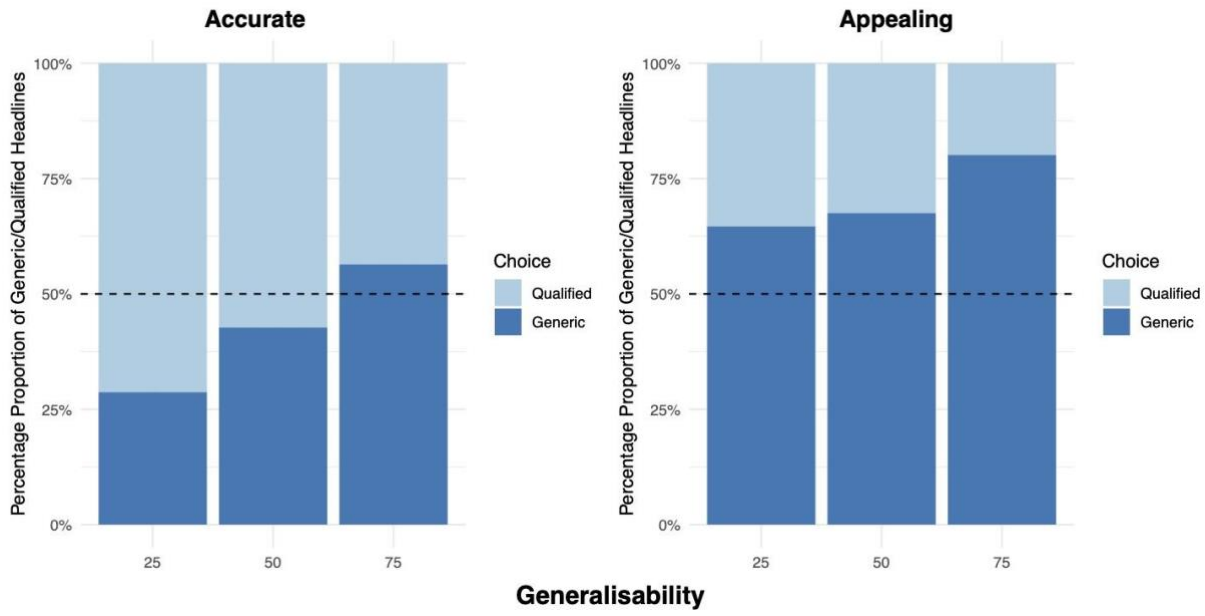
Table 10

Count, Expected Count and Adjusted Standardised Residuals of Generic and Qualified Headline choices, when data suggests 25%, 50% or 75% Generalisability across both levels of Motivation.

	Accurate			**Appealing**		
	Generalisability			**Generalisability**		
	25%	**50%**	**75%**	**25%**	**50%**	**75%**
	Generic Choice					
Count	56	85	110	128	131	157
Expected Count	83.10	84.80	83.10	140.08	137.25	138.67
Adjusted Std. Residual	-4.798	0.035	4.763	-2.317	-1.205	3.526
	Qualified Choice					
Count	139	114	85	70	63	39
Expected Count	111.90	114.20	111.90	57.92	56.75	57.33
Adjusted Std. Residual	4.798	-0.035	-4.763	2.317	1.205	-3.526

Figure 20

Stacked bar charts showing percentage proportion of headline choice (Generic, Qualified) in the 25%, 50% and 75% generalisability conditions, across the levels of Motivation (Accurate, Appealing).



Accurate

There was a significant association between Generalisability and Headline Choice $\chi^2(2, N = 589) = 30.573$, $p < .001$, V = .23. When Generalisability was 25%, the minority of participants chose the Generic headline (28.72%), significantly less frequently than the expected cell count ($\tilde{r}_{ij}$ = -4.80). When Generalisability was 50%, the minority of participants chose the Generic headline (42.71%). This frequency did not significantly deviate from expected cell counts. When Generalisability was 75%, the majority of participants chose the Generic headline (56.41%), significantly more frequently than the expected cell count ($\tilde{r}_{ij}$ = 4.76).

Appealing

There was a significant association between Generalisability and Headline Choice $\chi^2(2, N = 588) = 12.822$, $p = .002$, $V = .15$. The majority of participants chose the Generic headline when Generalisability was 25% (64.65%), however, this frequency was significantly lower than the expected cell count ($\tilde{r}_{ij} = -2.32$). The majority of participants also chose the Generic headline when Generalisability was 50% (67.53%). This frequency did not significantly deviate from expected cell counts. Finally, the majority of participants again chose the Generic headline when generalisability was 75% (80.10%), significantly more frequently than the expected cell count ($\tilde{r}_{ij} = 3.53$).

Open-Ended Headlines (Exploratory)

A total of 251 participants chose to use the optional text entry box to write their own alternative headline, this was 21.3% of our final sample. The number of participants who chose to write their own headline per-condition ranged from 35 to 50, these data are presented in Table 11.

Table 11

Optional headline usage (total participants in condition) across our Motivation x Generalisability conditions.

	25%	**50%**	**75%**
Accurate	39 (195)	46 (199)	50 (195)
Appealing	35 (198)	42 (194)	39 (196)

The alternative headlines written by participants fell into three distinct categories. In the first category, participants added a specific quantifier. Quantifying was achieved by using an explicit quantifier (e.g., “new supplement boosts immune system in *most* men”), specifying

a proportion (e.g., half of men, 1 in 2 men) or specifying the percentage provided in the scenario (e.g., 75% of men).

The second category of headlines was those that included a hedging term or device to caveat the headline. Crompton (1997) defines a hedge as "*an item of language which a speaker uses to explicitly qualify his/her lack of commitment to the truth of a proposition*" (p. 281). Hedging was achieved by using an additional qualifying term (e.g., "new supplement *may* boost immune system function in men") or by making reference to the fact that a study was conducted (e.g., "*Study has found* new supplement boosts immune system function in men") rather than making a universal claim.

The third category of headlines identified was those written to be explicitly appealing. Appeal was achieved by using optimistic language. For example, describing the supplement as 'promising' (e.g., "new supplement shows promising results for immune system function in men"), or 'successful' (e.g., "new supplement is successful at boosting immune system function in men") among other terms such as 'proven', 'encouraging' and 'groundbreaking'.

Finally, a number of headlines were uncategorised. These were headlines that paraphrased the original headline without making any meaningful changes to the content, or extracts that did not provide a usable headline. The usage frequency of each category is presented in Table 12.

Table 12

Number of headlines (out of 251) that included (1) a quantifier, (2) a hedging device, (3) explicitly appealing language, and (4) a paraphrased (or uncategorised) headline.

Quantifiers	137
Hedges	78
Explicit Appeal	42
Uncategorised	37

Note - some headlines fitted into more than one category, so the sum exceeds 251.

6.1.6. Discussion

The aim of this experiment was to extend the findings of Haigh et al. (2022). As predicted, we found that across all three levels of generalisability, a significantly greater proportion of participants chose the generic headline when motivated to be appealing. We also found that across both levels of motivation, we observed a significant increase in the proportion of participants choosing the generic headline as the generalisability of data increased. In other words, the production of a generic statement was more likely when participants could state their claim based on data that was higher in generalisability. Contrary to our predictions, there was an association between generalisability and generic production (i.e., higher generalisability corresponded with higher rates of generic choice), when the primary motivation was *both* accuracy and appeal. We expected that we would only observe an association between generalisability and generic choice when participants are motivated by accuracy, so this finding is not in line with our predictions. Instead, the association between generalisability and generic choice was 'dampened' by a motivation for appeal. In the accurate condition the proportion of participants choosing the generic headline increased by 14.01% and 13.7% points as generalisability increased. In the appealing condition these increases were

2.93% and 12.57%, so whilst we observed associations in both levels of motivation, the effects are less pronounced in the appeal condition.

Notably, we found that when motivated to choose the most accurate headline, the majority of participants chose the generic headline only when the generalisability of the data was high (75%). On the other hand, when motivated to choose the most appealing headline the majority of participants chose the generic headline even when the generalisability of the data was very low (at 25%, 50% and 75%). For readers, this could mean that generic language does not imply a high degree of generalisability when the writer is motivated by appeal rather than accuracy.

Our participants were given the option to write their own headline. We found that a relatively small proportion (21.3%) of participants chose to utilise this option. When used, participants often chose to quantify the headline (e.g., 75% of men, most men), as well as choosing to hedge (or qualify) the headline (e.g., supplement *may* be effective, supplement improves immune system *in study*). In these cases it appears that participants are attempting to write a more accurate headline. Finally, some participants chose to make their headline explicitly more appealing than the fixed choice headlines. Here, we consider the implications of these key findings.

Collectively, our results here shed light on the many contextual factors that can influence the production of generic language. Whilst there now exists a large body of evidence supporting the proposition that generics can be used by writers to intentionally or unintentionally inflate the generalisability of a claim (Cimpian et al., 2010; DeJesus et al., 2019; Gelman & Roberts, 2017; Gutiérrez & Rogoff, 2003; Leslie, 2007; Leslie 2008; Rogoff, 2003), the current findings, alongside that of Haigh et al. (2022), evidences that the production of a generic claim is contingent upon the true generalisability of data (i.e., the evidence-base

for such a claim). Our findings here are also in agreement with Haigh et al. (2022) with respect to the generalisability 'boundary' observed in their data. The authors discuss that the majority of participants reliably chose the generic headline when scientific consensus reached 70% or more, a finding that is supported here since the majority of participants chose the generic headline at 75% generalisability, irrespective of whether they were motivated to be accurate or appealing.

This requirement for writers to observe a high level of generalisability may speak to the innocence of their intentions when generics are used to summarise research findings. DeJesus et al. (2019) report that among a large body of psychological journal articles (>1000 papers), more than 90% of codable articles were titled with generic language. These data follow a recent call for all scientific research papers to include a 'constraints on generality' disclosure to protect against a reader inferring the largest possible generalisations (Simons et al., 2017). Importantly though, whilst the landscape of research dissemination has become increasingly receptive to bold universal claims, such as shorter distribution formats and an increased likelihood of positive media attention (Dumas-Mallet et al., 2018; Weinstein & Sumeracki, 2017), it is clear based on our findings that authors of research papers are unlikely to adopt the use of generics solely to boost the success of a publication. Instead, the decision to title using a generic may be motivated, at-least in part, by data that is generalisable enough to justify their use. This is because the preference for the generic headline in this experiment (as in Haigh et al. 2022) increased as a function of increased efficacy (i.e., the greater the % of men seeing an improved immune system, the greater the proportion of participants who chose to headline the medical information article with a generic). Still, prior to this study there existed no generic language production experiments that explicitly manipulated a person's intentions.

One important implication of our findings is that generic production may be largely dependent on a person's motivation. The effects of our motivation manipulation demonstrate that at every level of generalisability, a motivation to be appealing (rather than accurate) significantly increased the proportion of participants choosing the generic headline. For example, consider participants in the 25% generalisability condition (i.e., 25% of men improved their immune system with the new supplement). When motivated to choose the headline that is 'most accurate', 28.7% of participants chose the generic option. By contrast, when motivated to choose the headline that is 'most appealing', 64.6% of participants chose the generic option. Comparable effects were also found at 50% generalisability (42.71% vs. 67.53%) and 75% generalisability (56.41% vs. 80.10%). These results are enlightening as they suggest that when a writer's intentions are to communicate accurately, the preference for a generic statement is relatively low. However, when explicitly motivated to report the data in an appealing way, the preference for a generic statement greatly increases.

Another implication of our findings is that the effects of motivation and generalisability appear to be independent of one another. The effects of motivation that we observed did not differ across all three levels of generalisability. In other words, irrespective of how effective the new supplement actually is, being motivated by appeal amplifies the preference for generic language to a similar extent. In practice, this could mean that an ill-intentioned writer can rationalise making a generic claim even in the absence of solid data supporting the claim.

Importantly though, we evidence here as an extension of existing research, that the likelihood of a writer producing a generic statement is influenced by intentions, as well as their means.

The use of generics within the media is problematic, particularly when the reports cover health research. Despite their widespread adoption by authors of journal articles (DeJesus et al., 2019), the fact that generics can be used to exaggerate the importance and generalisability of findings is less damaging in this context than in the media, because the typical reader base of a journal article (i.e., academics) are trained in research methods and implicitly understand the limitations of a sample. By contrast, media articles are predominantly read by the general public who are not trained in research methods, and are potentially more susceptible to the effects of generics. For instance, some readers will attribute generic statements attached to 'doctors' or 'experts' as meaning *all* doctors or experts (Haigh et al., 2020), and in the absence of an obvious counterargument (such as when reading news articles) the public will assume as default that scientific consensus has been achieved to a high degree (Aklin & Urpelainen, 2014).

Tackling generic language in the media is an inevitably difficult task, especially when a writer may be motivated by appeal. Unlike academics, whereby accuracy and integrity is a professional requirement, the news media will often set out to promote clicks, evident in the paucity of identifiably 'fake news' surrounding politics and science (Lazer et al., 2018). In an environment that is conducive to bold universal claims about research (Dumas-Mallet et al., 2018), it is not immediately obvious how to motivate journalists more by accuracy than appeal. However, the issue could be remedied by continuing to improve the reporting standards of newly published research. Sumner et al. (2014) have evidenced that exaggerations in university press releases are associated with comparable exaggerations in the news media, suggesting that reductions in universal claims made by original authors could inspire a reduction in universal claims reported by the media. As such, authors must be consciously

aware of the potentially misleading nature of generics, and adopt a method of summarising their research that is accurate yet considerate of the limitations of their sample.

A number of methodological strengths give merit to the findings of this experiment. Firstly, our scenario is played out within the context of health communication. This means that our findings are widely applicable to the everyday individual, as health research is perhaps most frequently reported in the media (Bossema et al., 2019; Bratton et al., 2020). Secondly, our scenario is reflective of the communicative process from original data source into secondary reporting. Understanding generics in this context is important because, as discussed, the implications of generics are likely more consequential when read by a lay audience (i.e., readers of news articles). Moreover, a scenario that places the participant in the position of a journalist (rather than an academic) is a more effective way to manipulate motivation (i.e., asking an academic to select the 'most appealing' headline appears counterintuitive).

It is also important to consider the limitations of this experiment. Initially, one could argue that our forced choice measure is too restrictive. In other words, one may question whether fewer participants would choose the generic headline, had they been given more choices (e.g., 'many men'). However, the relative unpopularity of our open-ended question suggests that this is unlikely. Only a fifth of our sample decided to devise their own alternative headline, indicating that the options that we presented in this experiment were adequate for headlining the article.

Given that our scenario is hypothetical the ecological validity of our study is limited. Attempting to simulate genuine language production is unachievable without relinquishing a large amount of experimental control, meaning that we cannot make literal inferences based on these data. Also, whilst our scenario presents participants with fictitious data evidencing a positive change to public health, it is yet to be confirmed whether our key findings would also

be observed for negative changes to public health. For example, if participants were asked to choose the headline for an article that reports a weakened immune system from a new supplement, this may alter the decision-making process, given the general psychological principle that 'bad is stronger than good' (Baumeister et al., 2001). Future research should consider manipulating motivation and generalisability using scenarios that depict negative health outcomes, to establish the extent to which a re-contextualised outcome could amplify or dampen the effects observed in this study.

Of the research to date that has sought to understand the psychological mechanisms underpinning generic language, a large majority has focussed on how they are interpreted by their reader. This experiment aimed to extend the findings of Haigh et al. (2022) and understand what can influence a writer to actually produce a generic statement. A sample of English-speaking adults read a hypothetical scenario and chose either a generic headline (e.g., X improves Y in men) or a qualified non-generic headline (e.g., X improves Y in *some* men). We found a positive relationship between the generalisability of research findings and the proportion of participants choosing the generic statement. In other words, better data coincided with greater preference for the generic. We also found that a significantly greater proportion of participants chose the generic headline when motivated by appeal rather than accuracy. These findings suggest that a writer's decision to favour generic language is largely influenced by their explicit motivations, as well as their means to make a generic claim (i.e., how generalisable their data is). Continuing to improve academic reporting standards that achieve accuracy, whilst adequately addressing sampling limitations may help to reduce the number of overgeneralisations in the news media.

Author Contribution Statement

Study Conception: Harry Clelland, Matthew Haigh

Study Design: Harry Clelland, Matthew Haigh

Development of Materials: Harry Clelland

Pre-Registration: Harry Clelland

Data Collection: Harry Clelland

Data Analysis: Harry Clelland

First Draft: Harry Clelland

Review and Editing: Harry Clelland, Matthew Haigh

6.2. Generic Language in the Communication of Health Research (Exp. 8)

6.2.1. Abstract

Generic claims imply universal rules about categories or groups by glossing over individual variability and exceptions. Consider, for example, the generic news headline ‘Exercise helps teens quit smoking’ relative to the non-generic ‘Exercise helps *some* teens quit smoking’. When generic language is used in primary reporting of scientific research (i.e., journal articles) the reported claims are perceived to be more important and more generalisable than when non-generic language is used. This study aimed to establish whether the same effects would be present within secondary reporting of research (i.e., news articles aimed at the public), with a particular focus on personal health claims. Participants read a series of genuine news headlines in either their original generic format (e.g., ‘Exercise helps teens...’) a past-tense non-generic format (e.g., ‘Exercise help*ed* teens...’) or a qualified non-generic format (e.g., ‘Exercise helps *some* teens...’). Generic headlines were rated as slightly more important and more generalisable than headlines qualified with the word ‘*some*’. In contrast, we found no differences in perceived importance or generalisability between generic headlines and past-tense non-generic headlines. Our results suggest that writers must be explicitly non-universal when summarising research in order to accurately communicate constraints on generality.

Keywords: Generics; Language; Inference; Pragmatics; Generality; Science Communication

6.2.2. Introduction

Generic statements are utterances that express generalisations about categories (Cimpian et al., 2010). Take for example statement (1):

(1) Women give vaginal birth

The generic statement (1) expresses a characteristic (giving vaginal birth) on the basis of a group identity (women) and few people would argue the truthfulness of the statement. This example highlights one of the most distinctive characteristics of generic statements, which is that they have a tolerance to exceptions (Krifka et al., 1995). In other words, they are difficult to prove false. For example, consider generic (1) in comparison to its universal equivalent (2):

(2) *All* women give vaginal birth

The generic statement (1) can be asserted truthfully even in the face of obvious exceptions, such as the common knowledge that some women do not give birth, and some women give birth via cesarean section (roughly 1 in 4 women; NHS, 2019). By contrast, the non-generic universal statement (2) cannot be true in the face of exceptions, as it can be falsified with a single counterexample. A single woman who gives birth by cesarean section is sufficient to conclude that 'not all women give vaginal birth' (Lazaridou-Chatzigoga, 2019).

Another notable characteristic of generic statements is that they require minimal evidence to be accepted as true. Across four experiments, Cimpian et al. (2010) compared interpretations of generic statements about a fictional creature (e.g., "Lorches have purple feathers") to a quantified version of that statement, modified with the term 'most' (e.g., "Most Lorches have purple feathers"). Their results showed that despite knowledge of a low relative prevalence (i.e., less than 50% of Lorches have purple feathers), participants often judged the

generic statement to be true. Moreover, when asked to judge prevalence based on the generic statement, it was inferred that the property occurs in over 90% of category member cases (i.e., over 90% of Lorches have purple feathers). In comparison, no such finding was present for the quantified statement using '*most*'. In short, generics have three notable characteristics: they require minimal evidence to be accepted as true, they are difficult to prove false and they may imply universal rules on the basis of not-so-universal evidence.

One particular area of concern with regards to generalisability, relates to how scientific findings are communicated (e.g., journal articles, university press releases, social media posts). The findings of psychological research, which is based on a sample of humans, tends to be generically generalised to 'humans' or 'people', despite Henrich et al. (2010a) highlighting an over-reliance on 'WEIRD' samples (Western, Educated, Industrialised, Rich, Democratic) that are not necessarily representative of the global population. Additionally, the issue can be compounded by the process-of-translation from the originally authored journal article (primary reporting) to secondary reports of journal article findings written by a third party. Shorter distribution formats (e.g., blogs, Twitter) increasingly encourage researchers to make daring and 'snappy' claims about the wider applicability of their findings (Weinstein & Sumeracki, 2017). Similarly, exaggerations of this nature present in university press releases translate to news media reporting (Sumner et al., 2014), causing some researchers to call for explicit 'constraints on generality' disclosures in all scientific papers (Simons et al., 2017). Taken together, it would appear that there are increasing pressures for the nature of scientific communication to favour appeal in place of precision.

In the context of scientific writing, there is growing evidence suggesting that generic statements can be used to unwittingly imply that research findings are both more important and more generalisable than what can truly be inferred from the data (Cimpian et al., 2010;

DeJesus et al., 2019; Leslie, 2008). One way that they achieve this is by glossing over exceptions and variability on an individual level (Leslie, 2007). For example, a person may assert that "*men have short hair*" despite the wide individual variability in men's hair length. This may lead a reader to believe that the statement is highly generalisable. Furthermore, as discussed above, generics are resistant to counterarguments. For example, saying "*men have short hair*" is not disproved by the existence of a single man with long hair. This may lead a reader to believe that the implications of a generic statement are more important than a non-generic statement. Surprisingly, very few studies to date have tested the interpretation of generics in the context of scientific reporting.

To address the need for work in this area, DeJesus et al. (2019) conducted a series of experiments. First, they evaluated the prevalence of bare generic statements within the three most-read sections of academic journal articles (Title, Highlights, Abstract; Pain, 2016). They found that ~90% of codable papers titled their results with a generic statement (e.g., '*group discussion improves lie detection*'). There was also an overwhelming tendency for authors to make generic conclusions via limited samples. Through these findings, the authors conclude that typical language practice (i.e., generics) cause a disparity between the inherent limitations of research findings and the generality of the conclusions drawn from these findings.

Following this, DeJesus et al. (2019) experimentally manipulated genuine journal article titles to test how changes to their generic form would be interpreted by readers. Specifically, 'bare generic' headlines (e.g., group discussion improves lie detection) were altered to create multiple non-generic versions, such as a past-tense non-generic version (e.g., group discussion improv*ed* lie detection), a non-generic version qualified with the term 'some' (e.g., *some* group discussion improved lie detection) or a multi-cued non-generic version (e.g., *some* group discussion improved lie detection *under certain circumstances*). Overall, the

results of multiple experiments revealed that generic statements were rated as slightly more important than their non-generic counterparts. Moreover, generic statements were perceived to be more generalisable and more conclusive than non-generics, but only under particular circumstances.

The work of DeJesus and colleagues (2019) is enlightening as it highlights a seemingly discipline-wide tendency to over-generalise findings by making generic claims in article titles and abstracts. Their work was limited to primary source journal articles, written for research professionals. In the study below we aim to extend the findings of DeJesus et al. (2019) by investigating how generics are interpreted when used by the media to report research to the public. We focus specifically on the communication of health-related research, as this field of research is frequently reported by the media (Bossema et al., 2019; Schat et al., 2018) and may inform consequential health-related decisions by members of the public.

6.2.3. The Present Research

The aim of the current study was to extend the findings of DeJesus et al. (2019) and test whether generic statements would be interpreted as more important and more generalisable than non-generic statements when reported through a secondary source (i.e., the headline of a news article). The reason for our decision is two-fold.

First, this approach has wider applicability because a lay person is more likely to read news headlines that present short summaries of research than they are to read a primary source. In this study, headlines were rated by the same audience that will read them in the real world (i.e., the public), which was not the case in DeJesus et al. (2019), as non-academic participants evaluated the importance and generalisability of research papers written for an academic audience. This distinction has important implications because academics (via their

research methods training) have an implicit understanding of the limitations of a sample (i.e., they are aware that a generic claim does not apply across-the-board). This is not necessarily true of the public however, meaning that overgeneralisations caused by generics may have greater consequences in this context.

Second, generic claims may be more widespread in the media, as journalists are more motivated to utilise them. Generics are inherently short, simple and easy to understand, all of which are desirable properties for news article writers. While academics should fulfil their obligation to rigour and factual accuracy, secondary reporters are tasked primarily with translating the findings to a wider audience in an appealing manner. For this reason, generics may be more widespread in secondary reports because journalists are more motivated by appeal (and less motivated by precision) than the original authors.

We chose to focus specifically on reports of health research. This is one of the most common types of research covered by the media (Bossema et al., 2019; Bratton et al., 2020; Buhse et al., 2018; Schat et al., 2018; Sumner et al., 2016), largely because human health is relevant to everyone. Perhaps more importantly, we focused on media reports of health research because these reports could inform consequential health decisions made by the public. If research findings presented in a generic manner are perceived as more important and more generalisable, then individuals may be more likely to adopt associated lifestyle changes (for better or worse). There is therefore a potential danger that generic summaries of nuanced health research will cause some people to make health decisions based on the ill-informed assumption that research findings apply across-the-board.

The experiment reported below examines whether the use of generic phrases in the secondary reporting of health research increases the perceived importance and generalisability of findings. We selected 18 genuine news headlines. The headlines made a generic claim

about a group of people and summarised a piece of primary research, showing improvements in an aspect of health (e.g., '*online therapy helps teens recover from chronic fatigue syndrome*'). For each headline we created a past-tense non-generic version (e.g., '*online therapy helped teens recover from chronic fatigue syndrome*') and a non-generic version qualified with the term 'some' (e.g., '*online therapy helps some teens recover from chronic fatigue syndrome*'). Participants rated each headline on four measures. They firstly rated the importance of the research, and they then rated the generalisability of the research across three measures (sample diversity, sample size, universality of findings).

We expected that the effects observed by DeJesus et al (2019) data relating to summaries of primary data would carry over to secondary reporting of data. We hypothesised therefore that bare generic headlines (i.e., X helps Y) will be perceived as more important and more generalisable than both past-tense non-generic headlines (i.e., X helped Y) and qualified non-generic headlines (i.e., X helped some Y).

Overview of Method

This study was pre-registered prior to data collection. Materials, raw data and the analysis script can be found on the Open Science Framework (OSF; see osf.io/krpab/). The aim of the experiment was to extend the findings of DeJesus et al. (2019) by examining how generic phrases used in the secondary reporting of health-related research findings are interpreted by readers. Participants were shown genuine news headlines in three different formats. Each headline was presented to participants in either its original bare generic form (e.g., Plasma therapy helps critically ill Covid-19 patients), a past-tense non-generic (e.g., Plasma therapy help*ed* critically ill Covid-19 patients) or a non-generic qualified with the term '*some*' (e.g., Plasma therapy helps *some* critically ill Covid-19 patients). For each headline, we measured how important and generalisable (universal) the findings were perceived to be.

6.2.4. Method

Design

This experiment used a one-factor counterbalanced (Latin Square) repeated measures design. The independent variable (fixed factor) was the Headline Format with three levels (Bare Generic, Past-Tense Non-Generic, Qualified Non-Generic). Participants were randomly assigned to one of three Latin Square presentation lists, each of which contained 18 health research headlines presented in random order (see osf.io/6jgb7/ for Latin Square design). Participants were exposed to all three conditions (six bare headlines, six past-tense, six qualified), but saw only one version of each headline. After reading each headline, they answered four questions.

The first dependent variable measured how important the research finding was perceived to be (DV1). The final three dependent variables measured aspects of generalisability. We firstly measured the extent to which the research findings were thought to extend to people from diverse backgrounds (e.g., differing in race, ethnicity, socioeconomic status etc.) (DV2). Second, we asked participants to estimate the study sample size (DV3). Finally, we asked how generalisable the research findings were perceived to be (DV4).

Participants

We conducted a prospective power analysis for a one factor counterbalanced repeated measures design (one fixed effect, two random effects) with 18 items. To detect a 'medium' sized effect 80% of the time (d = 0.5, α = 0.05) we required a minimum of 69 participants (Westfall, Kenny & Judd, 2014). This was our minimum sample size target.

An initial sample of 102 consenting participants were recruited via Prolific online recruitment platform. We excluded one participant who did not complete the survey. All of the

remaining participants declared that they took the survey seriously. The final sample was made up of 101 English-speaking adults (38 male, 63 female), aged between 18 and 77 (mean age = 37.37, SD = 14.08). Of these, 34 were allocated to Item List 1, 34 were allocated to Item List 2, and 33 were allocated to Item List 3 (see osf.io/6jgb7/). Participants were paid £1. The power of our final sample (N=101) to detect a medium size effect (*d*=0.5) was d=0.83.

Materials

We selected 18 online news headlines that summarised research findings using a generic statement. These were sourced from the Google News archive, by searching for common terms (e.g., 'patients recover', 'helps women' etc.). Headlines were selected on the basis of two key characteristics. First, the headline must make a generic statement (e.g., '*pet dogs help kids feel less stressed'*, '*exercise helps men regain bone mass'*). Second, the headlines must summarise a piece of primary research that appears to show improvements in an aspect of personal health (e.g., '*online therapy helps teens recover from chronic fatigue syndrome*', '*music helps patients recover more quickly if played before surgery*').

Whilst our headlines were varied in content, each can be reduced to a generic 'X helps Y' format, whereby 'X' is a treatment or intervention that 'helps' (e.g., 'improves', 'boosts', 'beneficial to') a defined group of people 'Y' (e.g., 'patients', 'men', 'children'). One headline followed the inverse of this format by starting with the group (e.g., 'stroke patients [Y] recover arm use with virtual reality [X]') but served the same functional purpose as an 'X helps Y' statement. The headlines related to six categories of people (3 headlines per category; patients, children, women, men, elderly, teenagers). Full details of all 18 items (as well as their past-tense and qualified versions) including links to the original news articles can be found on the OSF (see osf.io/3mav5/).

Headline Format Manipulation

Each bare generic headline was manipulated to create two distinct non-generic versions. In its original wording, each headline is a bare generic statement. First, we created a past-tense non-generic version. For example, the generic "*music app helps Alzheimer's patients*" became "*music app helped Alzheimer's patients*". Second, we created a non-generic version that was qualified with the word '*some*'. For example, the generic "*music app helps Alzheimer's patients*" became "*music app helps some Alzheimer's patients*". Examples of our manipulation are presented in Table 13.

Table 13

Three Examples of the Bare Generic, Past-Tense and Qualified Headlines used in the Experiment.

a) "Stroke Patients Recover Arm Use With Virtual Reality"	
Bare Generic	*Stroke Patients Recover Arm Use With Virtual Reality*
Past-Tense	*Stroke Patients **Recovered** Arm Use With Virtual Reality*
Qualified	***Some** Stroke Patients Recover Arm Use With Virtual Reality*
b) "Plasma Therapy Helps Critically Ill Covid-19 Patients"	
Bare Generic	*Plasma Therapy Helps Critically Ill Covid-19 Patients*
Past-Tense	*Plasma Therapy **Helped** Critically Ill Covid-19 Patients*
Qualified	*Plasma Therapy Helps **Some** Critically Ill Covid-19 Patients*
c) "Testosterone Therapy Beneficial to Men with Heart Disease"	
Bare Generic	*Testosterone Therapy Beneficial to Men with Heart Disease*
Past-Tense	*Testosterone Therapy **was Beneficial** to Men with Heart Disease*
Qualified	*Testosterone Therapy Beneficial to **Some** Men with Heart Disease*

Measures

We based our measures on those used by DeJesus et al. (2019) study 2. We constructed our questions with the intention of replicating fundamental elements of the wording used by DeJesus et al., whilst adapting the formatting to make each question specific to our health-related headlines. For example, when measuring generalisability participants in the DeJesus paper were asked '*what percentage of those in the world today would show the effect described*'. These instructions were universal and applied to 60 summaries in total. By contrast, we made this question specific to each summary. Taking an example from Table 13, we measured generalisability by asking 'what percentage of *stroke patients* would *recover arm use with virtual reality*'.

Importance (DV1)

Subjective ratings of importance were established by asking participants how important they thought the research findings were, based on their interpretation of the headline. For each item, participants were asked to '*Please give your best guess of how* ***important*** *this finding is*'. Importance was rated on a scale ranging from 1 (Not at all Important) to 7 (Extremely Important).

Diversity (DV2)

Subjective ratings of diversity were established by asking participants the extent to which they believed the research findings would extend to those from diverse backgrounds, based on their interpretation of the headline. For all items, participants were given additional information on standard sampling procedures for the research being summarised. They were told that '*Every project is based on a sample of participants. Each sample includes a set*

number of people who completed the research project and each sample has unique characteristics.'.

The wording of the question was specific to the headline being interpreted. For headline 2 for example (see Table 1), summarising research on Covid-19 patients, participants were asked '*In your opinion… How* ***likely*** *is it that the effect described in the research project would* ***extend to Covid-19 patients from diverse backgrounds*** *(for example, differing in nationality, race, ethnicity, socioeconomic status, etc.)?'.* The subject of the question was modified based on the item (e.g., a summary about women was modified to '***extend to women from diverse backgrounds***'). Diversity was rated in likelihood on a scale ranging from 1 (Not at all Likely) to 7 (Extremely Likely).

Sample Size (DV3)

Subjective ratings of sample size were established by asking participants to estimate how many people took part in the research, based on their interpretation of the headline. For each item, participants were asked to *'Please give your best guess of* ***how many people*** *participated in this research project from the options provided.'.* As in DeJesus et al. (2019), estimated sample size was chosen from 1 of 7 options (1-10 people, 11-50 people, 51-100 people, 101-250 people, 251-500 people, 501-1000 people, 1001+ people).

Generalisability (DV4)

Subjective ratings of generalisability were established by asking participants to estimate the percentage of those described in the headline (e.g., teens, Alzheimer's patients etc.) that would show the effect described (e.g., helped to quit smoking, recover from surgery).

The wording of the question was specific to the headline being interpreted. In reference to headline 2 once more, which describes that plasma therapy helps critically ill Covid-19

patients, participants were asked *'Please give your best guess of what* ***percentage of critically ill Covid-19 patients*** *in the world today would be* ***helped by plasma therapy****.'* Generalisability was rated as a percentage ranging from 0 to 100%.

Procedure

This online study was run using Qualtrics survey platform (Snow & Mann, 2013). Participants accessed the study anonymously via the Prolific recruitment platform. They provided informed consent, before being given instructions that read as follows:

"In this survey, you will see a series of brief summaries of different health-related research projects. These summaries are based on the titles of genuine media articles. There are 18 summaries in this survey. You might see words that you do not recognise or jargon. Please give your first impression of each summary. Read the summary once or twice and give the first answer that comes to mind."

They were then randomly allocated to one of three item lists, detailed in our Latin Square design (see osf.io/6jgb7/). Once allocated, a series of 18 headlines were shown in a different random order to each participant. For each headline, participants read either a bare generic, past-tense non-generic or qualified non-generic version of that headline. They then indicated their interpretations of importance, diversity, sample size and generalisability. Afterwards, participants were asked whether they took the survey seriously, and whether or not we should use their data (Aust et al., 2013). Those who indicated a non-serious response were still paid for their contribution. The end-of-survey message included debrief information as well as a link to return to Prolific. Median completion time for the entire survey was 8.12 minutes. A .qsf file for the Qualtrics survey, which allows for direct replication of the study, can be accessed via the OSF (see osf.io/azjmp/).

6.2.5. Results

The following analysis was pre-registered prior to data collection. The pre-registration, as well as raw data and analysis script can be found on the Open Science Framework (see osf.io/wgz7d for pre-registration, and osf.io/krpab/ for data and script).

All questions required a response, so there was no missing data. No data points were excluded. Analyses were conducted using R version 4.0.2 (R Core Team, 2020) and RStudio version 1.3.959 (RStudio Team, 2020). Each dependent variable was analysed using a linear mixed effects model, fitted via the 'lme4' package (Bates et al., 2014). The fixed factor was Headline Format which had three levels (bare generic vs. past-tense vs. qualified). We obtained *p* values for our fixed effects using 'lmerTest' (Kuznetsova, Brockhoff & Christensen, 2017), which uses Satterthwaite's method for approximating degrees of freedom. Participants and items were entered into the model as crossed random factors. Following the recommendations of Barr et al. (2013), each model began with a maximal random effects structure (i.e., random intercepts and slopes for participants and items). If the model did not converge, we systematically reduced the random effects structure until the model did converge. The first reduction method was removing the random correlations. Next, running a random slopes only model, followed by a random intercepts only model if still unsuccessful. For all of our measures, the most maximal converging model was a random intercepts only model (i.e., maximal model with random slopes removed). Post-hoc tests were conducted using the 'emmeans' package (Lenth, 2022) using Satterthwaite's method for approximating degrees of freedom. Condition means are displayed in Figure 21. Statistics from the linear mixed models are presented in Table 14.

To establish whether there was an influence of the group being referred to in the headline (patients, children, women, men, elderly, teenagers), we ran additional ANOVA's

with group included as a factor. Group did not significantly interact with headline format on ratings of importance $F(10, 1800) = 0.484$, $p = .901$, diversity $F(10, 1800) = 0.347$, $p = .968$, sample size $F(10, 1800) = 0.304$, $p = .980$ and generalisability $F(10, 1800) = 0.567$, $p = .842$.

Figure 21

Density plots showing (1) mean importance rating, (2) mean diversity rating, (3) mean sample size estimates, and (4) mean generalisability estimates for each headline type (Bare Generic, Past-Tense Non-Generic, Qualified Non-Generic). Black lines represent condition means. Surrounding bands represent 95% CI's. Black dots represent raw data points, and beans represent smoothed density curves (N = 101).

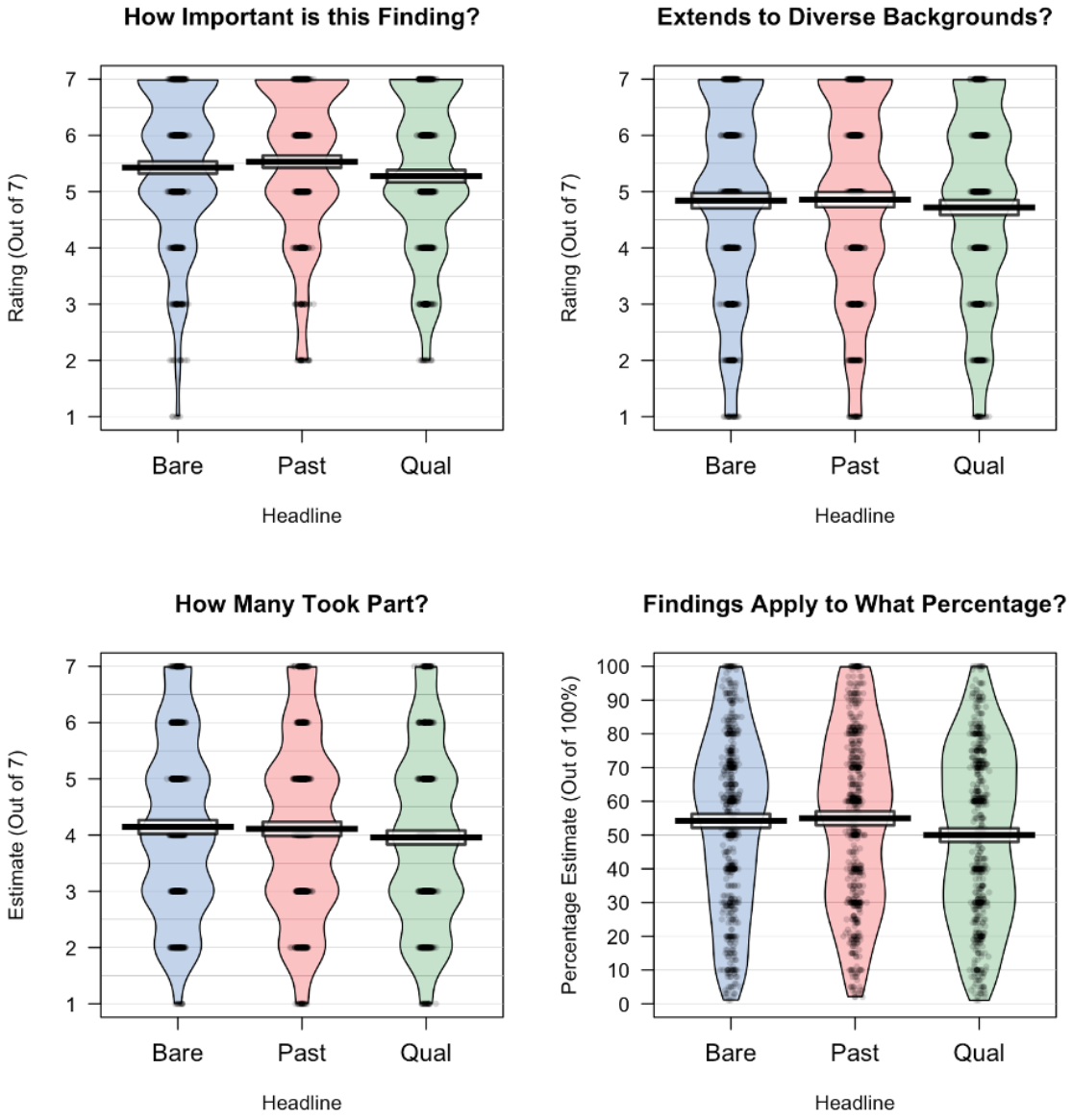


Table 14

Parameter estimates for (1) Importance, (2) Diversity, (3) Sample Size, and (4) Generalisability for each headline format (Bare vs. Past-Tense vs. Qualified; N = 101).

Condition	*b*	*SE*	*t*	*p*
Importance				
Bare (Intercept)	5.430	0.150	36.302	< .001
Past-Tense	0.105	0.060	1.750	.080
Qualified	-0.157	0.060	-2.618	.009
Diversity				
Bare (Intercept)	4.841	0.149	32.518	< .001
Past-Tense	0.017	0.067	0.250	.802
Qualified	-0.124	0.067	-1.855	.064
Sample Size				
Bare (Intercept)	4.145	0.152	27.350	< .001
Past-Tense	-0.035	0.061	-0.569	.569
Qualified	-0.188	0.061	-3.092	.002
Generalisability				
Bare (Intercept)	54.225	2.446	22.172	< .001
Past-Tense	0.738	1.089	0.677	.498
Qualified	-4.243	1.089	-3.895	< .001

Importance

We found a significant overall model, whereby headline format significantly predicted importance rating.

Bonferroni corrected post-hoc comparisons (using $\alpha = .0167$) revealed that Qualified headlines (mean = 5.27) were rated as significantly less important than Bare generic (mean = 5.43; $t(1700) = 2.618$, $p = .009$) and Past-Tense (mean = 5.54; $t(1700) = 4.368$, $p < .001$)

headlines. There was no difference between the Bare and Past-Tense conditions ($t(1700) = -1.750$, $p = .080$).

Diversity

We found a significant overall model, whereby headline format significantly predicted diversity rating. Bonferroni corrected post-hoc comparisons (using $\alpha = .0167$) revealed no significant differences between Bare (mean = 4.84) and Past-Tense (mean = 4.86; $t(1700) = -0.250$, $p = .802$), Bare and Qualified (mean = 4.72; $t(1700) = 1.855$, $p = .064$) and Past-Tense and Qualified conditions ($t(1700) = 2.106$, $p = .035$).

Sample Size

Participants gave sample size ratings on a 7-point Likert scale, with higher ratings indicating a larger sample size (1: 1-10 people, 2: 11-50 people, 3: 51-100 people, 4: 101-250 people, 5: 251-500 people, 6: 501-1000 people, 7: 1001 people or more).

We found a significant overall model, whereby headline format significantly predicted sample size estimates. Bonferroni corrected post-hoc comparisons (using $\alpha = .0167$) revealed that Qualified headlines (mean = 3.96) were estimated to have significantly less participants in the sample than Bare (mean = 4.14; $t(1700) = 3.092$, $p = .002$) and Past-Tense (mean = 4.11; $t(1700) = 2.523$, $p = .012$) headlines. There was no difference between the Bare and Past-Tense conditions ($t(1700) = 0.569$, $p = .569$).

Generalisability

We found a significant overall model, whereby headline format significantly predicted estimates of generalisability. Bonferroni corrected post-hoc comparisons (using $\alpha = .0167$) revealed that Qualified headlines (mean = 50.0) were rated as significantly less generalisable than Bare (mean = 54.2; $t(1700) = 3.895$, $p < .001$) and Past-Tense (mean = 55.0; $t(1700) =$

4.573, $p < .001$) headlines. There was no difference between the Bare and Past-Tense conditions ($t(1700) = -0.677$, $p = .499$).

6.2.6. Discussion

The aim of this study was to extend the findings of DeJesus et al. (2019). In a series of experiments, DeJesus et al. found that journal articles titled with generic statements were perceived to be more important than non-generic versions of the same title. They were also perceived as more generalisable and conclusive under particular circumstances. The current study aimed to extend these findings by identifying whether the observed effects carried over to secondary reporting of primary research findings (i.e., media headlines). We compared the interpretation of genuine generic headlines (i.e., X helps Y) to two non-generic versions; a past-tense non-generic version (i.e., X helped Y) and a qualified non-generic version (i.e., X helps some Y). We expected that generic media headlines would be rated as more important and more generalisable than the two non-generic versions.

The key findings of this experiment are two-fold. First, we found that bare generic headlines (e.g., '*exercise helps men regain bone mass*') were perceived as marginally more important and generalisable than qualified non-generic headlines (e.g., '*exercise helps some men regain bone mass'*). Second, we found that bare generic headlines were perceived as no more important and no more generalisable than a past-tense non-generic (e.g., '*exercise helped men regain bone mass'*). In other words, 'X helps Y' was perceived no differently to 'X helped Y'. What follows below is a more detailed consideration of both of these key findings.

As predicted, we found that headlines including generic statements were considered to summarise findings that were more important than the same headlines qualified with the word 'some'. In other words, 'X helps Y' was considered to be more important than 'X helps *some*

Y'. The same pattern of results was also found for estimates of sample size and ratings of percentage applicability. Generic headline summaries of research findings were judged to have been concluded from a larger sample of participants than the qualified version (i.e., a larger sample was used to conclude 'X helps Y' than the sample used to conclude 'X helps *some* Y'). Finally, the same pattern of results was found for estimates of percentage applicability. Generic headline summaries were perceived to be more applicable to the wider population than the qualified version (i.e., 'X helps Y in men' was considered to be more generalisable to the wider population of men than 'X helps *some* Y').

Taken together, these findings largely align with our understanding of generics to date, in that they can exaggerate claims and imply universal rules attached to groups (Cimpian et al., 2010; Gelman & Roberts, 2017; Gutiérrez & Rogoff, 2003; Leslie, 2007; Leslie 2008; Rogoff, 2003). Importantly though, these findings develop our understanding by evidencing that the effects of generic language that have been previously observed in primary reporting (i.e., journal articles), are also observed in secondary reports designed for public consumption. This may be consequential, as implying that a piece of personal health research is more important and more generalisable than it really is may directly impact the health beliefs and behaviours of the reader.

An increased demand for shorter, more concise distributions formats encourages authors of journal articles to make enticing claims about their findings (Weinstein & Sumeracki, 2017), and the most exaggerated (or enticing) claims are favoured by the news media (Sumner et al., 2014). Whilst DeJesus and colleagues (2019) effectively evidence a widespread tendency for authors to use generics in the shortest sections of articles (titles), and also demonstrate the over-importance and over-generalisability achieved by using generics; the present data (using genuine news media articles titled with generics) indicate that the same

effects carry over into interpretations of secondary reports (i.e., media headlines). This may be particularly concerning, because arguably, the dangers associated with generics are more pronounced within this context. Media headlines are predominantly read by the general public, who by-and-large are not trained in research methods like academics are. There may therefore be greater potential for exaggerated claims to be taken literally by lay readers, because generics exploit the tendency for some readers to conclude that a statement applies to *all* members of a category (e.g., all men) (Haigh et al., 2020; Leslie et al., 2011).

Despite significant findings, it must be acknowledged that in line with DeJesus et al. (2019), the effects that we observed were small. Considering our Importance variable for example, average importance ratings from Qualified to Generic conditions elevated from 5.27 to 5.43 (on a 7-point Likert scale). The magnitude of these effects are comparable to DeJesus and colleagues, who, for example, found a small but significant difference between the importance ratings of bare generic vs. non-generics (4.19 vs. 4.03), as well as the generalisability ratings of framed generics vs. non-generics (55.70% vs. 53.47%, p. 18374).

One explanation for the small effects observed within this study, however, could be due to the nature of our stimuli. The headlines that we chose evidenced improvements to an aspect of personal health, meaning that our headlines were inherently *positive*. Yet, given the robust psychological principle that 'bad is stronger than good' (Baumeister et al., 2001), it is possible that if our chosen generic headlines were to be *negative,* we could expect to see larger, more substantial effects. For instance, a negative generic like '*vaping harms children's learning*' could be considered more important than a positive generic like '*vaping helps children's learning*'.

Perhaps our most notable finding though, was that there was no difference between the generic headlines and the past-tense non-generic headlines on ratings of importance or

generalisability. In other words, across all of our measures, generic statements (i.e., X helps Y) produced the same ratings of importance and generalisability as the past-tense non-generic version (i.e., X helped Y). A potential explanation for this could be that both statements make a universal claim about Y and are therefore perceived in the same way. By contrast, the qualified statement 'X helps some Y' makes a non-universal claim about Y that is limited to a certain number of individuals, hence is interpreted differently.

This unexpected finding may have significant implications, as it suggests a requirement for writers to be explicitly non-universal when looking to accurately convey limits on generality. DeJesus et al. (2019) found that their participants were more sensitive to language when journal article titles had multiple cues to non-generality (p. 18375). For example, a title that is both qualified *and* past-tense (i.e., X *helped* Y, *under certain circumstances*). The authors conclude that in order to accurately communicate constraints on generality (acknowledging the limitations of the sample from which results are drawn from), writers may be required to take an explicit linguistic approach, rather than a reliance on subtle signals. Our results here support this interpretation. Despite the fact that a generic statement phrased in the past-tense no longer fits the definition of a 'generic', the evidence here implies that there are no meaningful differences in how they are interpreted. This aligns with the notion that a writer cannot successfully acknowledge sampling limitations by taking a non-explicit approach to language (i.e., only slightly altering the generic statement format). Rather, substantial linguistic changes to the generic formula must be made in order to avoid unintentional inaccuracies. This conclusion does not align with the recent work of Peters et al. (2022), which argues (based on the data of De Jesus et al. 2019) that the overgeneralisation effects of generics should be avoided by framing them in the past tense. Based on our current findings though, we reason here that replacing a generic universal (X helps Y) with a past-

tense universal (X helped Y) is not substantial enough to accurately consider the limitations of a sample. Writers must instead be explicitly non-universal to achieve this (e.g., by qualifying their statement with '*some*').

Despite evidence presented here indicating the potentially misleading effects of generics within the media, the solution is not immediately obvious. DeJesus et al. (2019) acknowledge that it would be unreasonable to expect generics within academia to disappear overnight, but by way of comparison, tackling generics within the media may be considerably tougher. Simons et al. (2017) have called for 'Constraints on Generality' statements to be included in each academic paper, which as a solution, would appear to be straightforward and easily implemented given that academics are motivated to fulfil their obligation to rigour and accuracy. Writers of news media articles, on the other hand, place greater importance on appeal (Dumas-Mallet et al., 2018), meaning that it will be considerably harder to 'break the mould'. Optimistically though, avoiding exaggeration in original papers may carry over into the media. Sumner et al. (2014) have demonstrated a relationship between exaggerations in university press releases and exaggerations in the media reports of these releases. It may therefore be possible to reduce the number of universal claims made in the media, by continuing to improve academic standards for research summaries.

In conclusion, generic statements (i.e., X helps Y) offer writers a short, simple and easy-to-understand mode of communication. However, by using them to communicate research findings the writer may intentionally or unintentionally misrepresent the importance and generalisability of findings. We sampled English-speaking adults to interpret genuine news media articles, reporting on the results of a research study evidencing improvements to an aspect of public health (e.g., '*pet dogs help kids feel less stressed'*, '*exercise helps men regain bone mass'*). Our results show that when media headlines are summarised with a

generic statement (i.e., X helps Y), the findings were considered to be slightly more important, based on larger samples and more generalisable than summaries qualified with the word 'some' (i.e., X helps *some* Y).

This extends previous research by demonstrating that the observed effects are present in both primary and secondary reporting of research. Contrary to previous studies, we found no differences between interpretations of generic statements (i.e., X helps Y) and past-tense non-generics (i.e., X helped Y). Both statements make a universal claim about a group and were interpreted in the same way by the participants in our sample.

Future research should measure how health-related generics are interpreted when headlines are negative, as it reflects more closely the behaviour of the public (i.e., 'googling' symptoms). Continuing to improve reporting standards of academic papers to accurately consider the limitations of a sample appears to be the most effective way to minimise the potential overgeneralisations associated with generic claims.

Author Contribution Statement

Study Conception: Harry Clelland, Matthew Haigh

Study Design: Harry Clelland, Matthew Haigh

Development of Materials: Harry Clelland

Pre-Registration: Harry Clelland

Data Collection: Harry Clelland

Data Analysis: Harry Clelland

First Draft: Harry Clelland

Review and Editing: Harry Clelland, Matthew Haigh

6.3. Chapter Summary

Two experiments are reported in chapter six, designed to answer our third research question: '*How do pragmatic factors influence the production and interpretation of generic language?*'. We addressed this question by firstly testing production. Our first experiment tested the extent to which 'quality-of-evidence' (or generalisability), as well as motivation, could influence a writer's production of a generic statement. Then, our second experiment tested whether a generic media headline would cause readers to consider the findings more valuable than non-generic versions of the same headline. Below we outline the ways in which our findings from these experiments directly answer our final research question (RQ3).

Writers were more likely to produce a generic statement (over a qualified non-generic) when the data supporting the statement was more efficacious. This reflects the fact that writers feel more 'justified' summarising in a way that implies universality, when the data supporting the claim are more universal in nature. Writers who were motivated to make their headline appealing (rather than accurate) reliably chose the generic headline more often, irrespective of data generalisability. With this finding we advance the current understanding of generic language, by extending that of Haigh et al. (2022) and demonstrating that the production of generic language is largely associated with their communicative intentions (as well as the strength of evidence supporting the claim). Universal statements, these are generics and past-tense versions (e.g., '*therapy helps young people*' and '*therapy helped young people*'), are considered by readers to be more important and more generalisable than non-universal qualified statements (e.g., '*therapy helps some young people*'). This conclusion is based on genuine secondary reports of health research within the media, which extends the work of DeJesus et al. (2019), who report similar results based on primary reports of psychological research (i.e., journal article titles).

Chapter 7. General Discussion

7.1. Overview

The use of ambiguous language can often cause a discrepancy between what is *said* in an utterance, and what is *meant* by an utterance. In this thesis, we aimed to understand the contextual (pragmatic) factors that influence (1) the *production* of linguistic ambiguity, and (2) the *comprehension* of linguistic ambiguity, whilst communicating about human health. To address our research questions (outlined in chapter two), we aimed to establish the contextual factors affecting the use and interpretation of three distinct forms of ambiguous language (uncertainty terms, quantifiers, generic language). By running experiments in pursuit of these aims, we applied the principles of experimental pragmatics to better understand the psychological mechanisms that underpin language.

What follows in sections 7.2, 7.3 and 7.4 is a summary and discussion of what was discovered within each of our experimental chapters, as well as the implications of these findings. These chapters reported eight pre-registered, open science experiments with a total sample of 4143 participants. Then, in section 7.5 a set of evidence-based recommendations are issued to those tasked with communicating about health. Section 7.6 provides a critical evaluation of the methods and evidence reported in this thesis, as well as considerations for the direction of future research. Finally, section 7.7 outlines the overall conclusions drawn from the studies conducted in completion of this programme.

7.2. Indirectness Can Achieve What Uncertainty Terms Cannot

Words like ‘possibly’, ‘maybe’ or ‘might’ are uncertainty terms that are non-referential. In other words, they do not denote a solid (or tangible) likelihood (Holtgraves, 2014). The use of these terms creates the potential for misunderstandings between speaker and hearer, because there are multiple potential motivations for their use (Bonnefon & Villejoubert, 2006). If you’re told by your doctor that “*maybe, you’ll suffer side effects*”, the term ‘maybe’ could have been used to express the doctor’s genuine uncertainty about the likelihood of suffering side effects. However, it could also be the case that your doctor believes you will certainly suffer side effects, and is using ‘maybe’ as a way to soften the damaging news. This latter interpretation (that the term is being used to soften bad news) becomes more likely as the severity of the outcome is increased. For instance, if instead, the statement from the doctor was that “*maybe, you’ll suffer serious side effects*”, the person interpreting this statement will consider this outcome to be more likely, and will be more likely to consider the use of ‘maybe’ as a tactic for softening bad news.

One key issue in the literature on uncertainty terms is a disproportionate focus on how they are *interpreted* within a range of contexts (Bhise et al., 2018; Bonnefon et al., 2011b; Lee & Pinker, 2010; Pinker et al., 2008). In chapter four we set out to better understand the factors that can influence a speaker to *produce* an uncertainty term, and following the recommendations of researchers, we observed (via a large two-part study) how certain pragmatic factors affected both the production *and* subsequent interpretation of language. In other words, we used data from a production experiment to also test interpretation, thereby observing the communicative dyad (i.e., from the perspective of both the speaker and the hearer).

We designed an experiment that builds on the work of Holgraves and Perdew (2016). They hypothesised that when a person is tasked with communicating uncertain negative information, increasing the social stakes (face-threat) of a situation would induce a greater use of uncertainty terms like 'possibly' or 'maybe', as a politeness tactic for managing face. To their surprise, the use of uncertainty terms throughout their data was infrequent (irrespective of the level of face-threat), and so as an alternative, the authors developed their own scale of indirectness. When analysing the data in this way, a significant effect of face-threat was discovered, whereby more threatening situations were managed by speakers through the use of subtle indirectness, rather than explicit uncertainty terms (e.g., possibly).

In our study, we strengthened the manipulation of face-threat by asking participants to break bad medical news to a patient. Now, rather than communicating (relatively) unimportant information under informal circumstances (e.g., telling a friend that they're drinking too much), participants were communicating high-stakes information (i.e., a diagnosis) in a formal manner (i.e., doctor-patient) to a potentially vulnerable recipient. We predicted that these strengthened vignettes would encourage the production of uncertainty terms as a tactful strategy for managing face.

Across a number of measures, our findings aligned with that of Holtgraves and Perdew (2016). Contrary to our predictions, the number of explicit hedging terms (e.g., possibly, maybe) did not change under high face-threat. Instead, our findings suggest that threat is managed via subtle indirectness, and this is supported by a number of our measures. First, the number of words used to communicate the news was significantly greater under high face-threat. Second, participants used significantly more dispreferred markers (e.g., 'I'm sorry to tell you', 'Unfortunately') under high face-threat. Given that an indirect speech act is that which deviates from perfect efficiency (Pinker et al., 2008), it seems to be the case that rather

than saying ‘possibly’ when they really mean ‘likely’, a speaker will manage face by using extra (unnecessary) words to ‘tip-toe’ around bad news. This is a particularly important finding because whilst we know that indirect language can be used to manage lower-stakes situations (see Pinker et al., 2008), we have demonstrated through direct experimental manipulation that face-work under high threat is still achieved with subtle indirectness. This reluctance to deliver bad news in a health context has been previously dubbed as the ‘MUM’ effect (Rosen & Tesser, 1970; Tesser et al., 1971), and a ‘stalling’ delivery style is present in some (but not all) doctors (Shaw et al., 2012). Based on this, establishing what can influence the individual differences in reluctance when breaking bad news would be a valuable next step for research.

Notably, it seems that the formulation of indirect speech is particularly difficult to achieve. Participants under high face-threat took significantly longer to break bad news, but only when the diagnosis was uncertain (rather than certain). From a hearer's perspective, engaging in politeness depletes mental resources, as extra processing power is needed to ‘figure out’ the speaker’s meaning (Bonnefon et al., 2011a). Here we show that the same mechanism activates when producing politeness as the speaker. That is, simultaneously balancing the desire to accurately communicate uncertainty, whilst actively trying to avoid harming the recipient, requires (at the very least) some additional processing time.

Finally, in the second stage of this two-part experiment, a new set of participants rated the language produced under the circumstances of part 1. Our experimental manipulation of certainty in part 1 caused changes to language that were detected by participants in part 2. For example, uncertain speakers communicated a less probable and less direct diagnosis. However, our manipulation of face-threat caused no detectable changes in the nature of

language comprehended in part 2. These results hold implications for the effectiveness of communicating bad news to patients, discussed in more detail later in this chapter (see 7.5).

7.3. Unlike '1-in-X' Ratios, The Perception of a Tactful Quantifier is Not Influenced by Severity

Quantifiers like 'some' or 'most' are characterised as non-referential in the same way that uncertainty terms are (Holtgraves, 2014). As a consequence, they possess the same potential to cause misunderstandings between two communicators. If you're told by your doctor "*some people never make a full recovery*", the doctor may have used 'some' because they are genuinely uncertain about whether you will make a full recovery. However, it's also possible that the doctor believes that you will certainly not make a full recovery, and has qualified the statement with 'some people' as a way to avoid harm.

In contrast with verbal quantifiers, '1-in-X' ratios refer to an exact probability. The likelihood represented by 1 in 5, for example, conveys the chance of an outcome that is solid and not objectively variable (20%). Despite this however, expressing risk in this format is ambiguous, because '1-in-X' ratios are sensitive to cognitive biases (Pighin et al., 2011; Pighin et al., 2015; Sirota & Juanchich, 2019). A person who is told that "*1 in 100 people catch this disease*", will consider the chance of this occurring to themselves to be more likely than if they are told "*5 in 500 people catch this disease*" (N-in-X*N ratio format), even though the communicated probability is identical (1%).

In chapter five of this thesis we tested the pragmatic factors that could influence both the production *and* comprehension of utterances that quantify risk. In our first experiment we evidenced that from the perspective of a speaker, proportional quantifiers like 'some' or 'many' are chosen specifically to manage particularly face-threatening situations. Precisely,

when threat is high a speaker will choose to represent risk with a lower proportion quantifier (i.e., saying ‘some patients’ rather than ‘many patients’) as a politeness tactic for achieving face-work. This exact mechanism also occurs for the production of probabilistic statements (e.g., somewhat unlikely; Sirota & Juanchich, 2015). More than this, we evidenced that exposure to a highly threatening set of circumstances (relatively speaking) increases a speaker's intention to communicate in the interest of managing face. Plausibly then, our findings suggest that a speaker will attempt to match the extent of their face-management tactics with how much they feel the tactics are required, based on the sensitivity of the situation.

Our second experiment looked at how hearers interpret the use of proportional quantifiers when they are used to qualify a face-threatening statement. We discovered that when the outcome attached to an utterance is severe (compared to non-severe), the use of quantifiers like ‘a few’, ‘some’, ‘many’ and ‘most’ are reliably considered a marker of politeness rather than genuine uncertainty. In other words, “*many people suffer deafness*” is seen as polite more often compared to “*many people suffer insomnia*”. This is in keeping with Bonnefon and Villejoubert (2006), who evidenced the same pattern of results for ‘*possible deafness*’ vs. ‘*possible insomnia*’. Unexpectedly however, we found that a severe outcome did not alter the perceived likelihood of the event occurring, when qualified with a proportional quantifier. For example, the likelihood of “*a few people suffer insomnia*” was the same as the likelihood of “*a few people suffer deafness*”, and this repeated for all quantifiers. This is not the case for uncertainty terms like ‘possible’, which may make the use of proportional quantifiers a more effective way to communicate health risk to patients.

In our third study we ran two experiments to understand how known cognitive biases (1-in-X bias, severity bias) impacted estimations of health risk, as well as subsequent health-

based decision making. We hypothesised that a high severity outcome would interact with the overestimation of risk produced by a '1-in-X' ratio. Both biases caused hearers to overestimate their risk of catching a disease, but they did so independently (i.e., the two biases did not interact). Worryingly, the extremity of overestimation was large across-the-board, with overestimation in the '1-in-X' condition reaching an average of 21 percentage points above genuine risk. Put into perspective, the disease that was given a 1 in 13 chance, which equates to a 7.69% likelihood, was considered to have approximately a 29% chance of *actually* happening. This is four times higher than the genuine likelihood, supporting the arguments of research to wholly remove the '1-in-X' ratio format from health communication (Zikmund-Fisher, 2011, 2014). Finally, results from our decision-making measures suggest that whilst the severity of an outcome affects decisions made about health, this is not the case for '1-in-X' ratios. This contradicts previous research (Sirota et al., 2018a; Sirota & Juanchich, 2019), and may suggest that a greater psychological 'weight' is placed on how serious the potential outcome is.

Taken together, the findings detailed in chapter five demonstrate that under face-threatening circumstances, a speaker is more motivated to engage in face-work and will alter the communicated probability expressed by a proportional quantifier (e.g., some) in order to do so. A hearer interpreting these quantifiers in severe situations considers their use to be an indication of politeness, rather than an expression of genuine uncertainty. However, unlike uncertainty terms (e.g., possible), a person interpreting a quantifier does not react by upscaling their view of likelihood. In other words, they don't consider a severe outcome to be more likely, even though they consider the speaker to be motivated by tactfulness. Given that they are objectively less ambiguous, one might argue for the use of '1-in-X' ratios as an alternative to proportional quantifiers. However, these ratios cause a hearer to greatly overestimate risk,

an effect that is compounded by a severe outcome. For this reason, a proportional quantifier may be more suitable for achieving accurate health communication whilst also attending to face.

7.4. Universal Headlines Are Appealing, But They Risk Overgeneralisation

The ambiguity associated with a generic statement is largely due to their resistance to exceptions (Krifka et al., 1995). Imagine that a researcher has run a study looking at the effects of an 8-week mindfulness intervention on smoking habits in men. In their analysis, the researcher discovers that of the 25 participants recruited, 3 participants displayed a reduction in smoking habits. Even though this is far from 'strong' data, the researcher could truthfully state (based on their results) that "*Mindfulness reduces smoking in men*". This is because exceptions to the rule do not falsify the statement, so even though '*not all*' men see a reduction in smoking, the generic statement can still be considered true (Lazaridou-Chatzigoga, 2019). Generics are therefore inviting for speakers (or writers) in communication about health, because they can be used to overstate the importance and generalisability of findings, irrespective of the data supporting the claim (Cimpian et al., 2010; DeJesus et al., 2019).

In chapter six of this thesis, we first aimed to understand the pragmatic factors that could influence a person to produce a generic statement. In our first experiment, we predicted that a writer would choose a generic statement (rather than a non-generic statement, qualified with the word 'some') more often when the data supporting the statement was stronger, but also when they were motivated by appeal (rather than accuracy). Our results confirmed both of these predictions. Writers chose to headline a news media article with a generic statement firstly when the data underlying the claim was more efficacious (e.g., 75% vs. 50%), and secondly when they were explicitly motivated to select the most appealing (rather than the most accurate) headline. This experiment serves as one of a limited number of generic

language production experiments, and demonstrates that whilst generics are produced more frequently the more they are justified by their evidence-base, it is also true that a writer who is motivated to generate a 'snappy' headline only requires minimal evidence to do so. To demonstrate this, when the generalisability of data was only 25% (i.e., effective in a quarter of people), those motivated by accuracy chose the generic ~29% of the time. By contrast, those motivated by appeal chose the generic ~64% of the time.

In our second experiment we observed the interpretation of generic statements, as it applied to news media headlines summarising health research. The aim here was to establish whether the exaggerated importance and generalisability associated with generically summarised psychological journal articles (DeJesus et al., 2019), would extend to news media headlines often consumed by the general public. We discovered that generic health media headlines are considered more important and more generalisable than their qualified non-generic counterpart (e.g., qualified with 'some'). However, there were no differences between the generic headlines and the past-tense non-generic headlines (e.g., Plasma therapy helps covid patients vs. Plasma therapy helped covid patients). The qualified non-generic headline versions were the only versions that do not make a universal statement about a category (e.g., covid patients), which suggests that in order to properly express constraints on generality, a writer must be explicitly non-universal in order to do so.

The conclusions drawn here differ from contemporary research. Based on the data of DeJesus and colleagues (2019), a recent paper by Peters et al. (2022) argues that research results should be reported in the past-tense as a way to avoid overgeneralisation. The results of our second experiment submit that this recommendation does not go far enough. Instead, it seems to be the case that writers must qualify their statement (e.g., Plasma therapy helps *some*

covid patients), because overgeneralisation only subsides when a statement is explicitly non-universal in this way.

Collectively, the experiments reported in chapter six show that generic statements are particularly appealing to writers. This is such that when a generic is used to summarise research in the media, there may not be sufficient data to justify the claims being made, especially when the author is motivated solely to garner the attention of the reader. In order to counter the untoward consequences of generic language (i.e., overstating research importance and generality), it is necessary for writers to prominently preface their summaries. Taking steps to improve the reporting standards of published research is likely an effective way to kickstart this process, particularly considering that exaggerations in primary reports of research (i.e., journal articles) often translate into exaggerations in the news media (Dumas-Mallet et al., 2018; Sumner et al., 2014).

7.5. Recommended Lines-of-Enquiry for Effective Health Communication

As discussed in chapter one, health professionals are faced with the difficult task of attending to the emotionality of a patient's circumstances, whilst simultaneously endeavouring to communicate with both accuracy and clarity under uncertainty. These tasks do not go hand-in-hand, because attending to face requires the formulation of politeness strategies, yet these strategies often involve effectively 'bending the truth' of a proposition. In other words, communicating in a way that converts the information *intended*, into altered information which is actually *expressed*. It is certainly the case that the most effective method of communicating with patients is one that comes closest to achieving both tact and accuracy.

Despite this, studies of doctor-patient communication highlight that not much is known about the specific elements of language that make this process effective (Burgers et al., 2012).

Papers that address precisely how a doctor should construct their language are rare (Paul et al., 2009), and of those that do, they did so from the point-of-view of general communicative 'styles' (e.g., comforting strategy; Sparks et al., 2007) rather than more concrete linguistic devices. In this section therefore, we attempt to partly address this issue, by considering which forms of communication appear (based on our data) to be the 'path of least resistance' when expressing health risk to patients. To do this, we outline below two propositions for health communicators. We consider these propositions to be the most fruitful avenues for future work, such as clinical research designed to test their potential effectiveness in healthcare settings.

(1) Aim to be subtly indirect when breaking bad news

Our first proposition (1) is based on a two-part experiment reported in chapter four. In the second part of this study, participants rated the language that was produced by a separate set of participants (acting as the speaker) in part one. We found that when diagnosing a patient with a serious condition (i.e., breaking bad news), there was no effect of face-threat on ratings of probability and certainty. Specifically, the language used to break bad news in part 1 communicated the same probability (i.e., likelihood that the patient has the condition) and the same certainty (i.e., doctors level of commitment to the statement) no matter the referent (i.e., whether the news was delivered directly to the patient or not). Importantly, this indicates that the politeness strategies used in part 1 did not alter the communicated probability or the perception of the doctor, in the way that politeness strategies typically do.

Promisingly then, considering that face-threat altered the language of participants in part 1 across a number of measures, the politeness strategies used to address face-threat seem to effectively achieve face-work whilst avoiding potential issues with increased ambiguity. To recap, those tasked with breaking bad news directly to the patient did not use more explicit

uncertainty terms (e.g., possibly, maybe). Instead, they used significantly more dispreferred markers (e.g., 'I'm sorry to tell you…') to be conciliatory towards the patient, and they used more overall words to deliver the news. Both of these strategies are consistent with a more subtle politeness strategy, characterised by indirectness. In short, we propose that speakers should seek to avoid overly direct language (e.g., 'you have Alzheimers'), because whilst this achieves clarity, it is often considered rude and insensitive (Blackhall et al., 2001), whereas indirectly delivering bad news has the added benefit of maximising face-management.

Indeed, previously published recommendations for doctors have advocated for indirectness. Barclay et al. (2007) note that indirectness may be preferable when delivering bad news to patients from threat-sensitive cultures, and the work of Sparks et al. (2007) reports case studies within which an indirect approach has led to patient's satisfied with their diagnosis. There is an important distinction to be made however, between indirectness and *subtle* indirectness. For instance, Kakai (2002) describes indirect communication as disclosing little to no information. This is reflective of the 'stalling' approach reported by Shaw et al. (2012), which (out of three delivery styles) resulted in the most patient confusion and the worst doctors scores for breaking bad news. Our proposition (1) advocates for a more subtle form of indirectness, an example of which being the use of dispreferred markers (e.g., "I'm sorry to have to tell you…"). The widely applied recommendations of Baile et al. (2000) stipulate that the doctor should address the patient's emotions with empathy, so the use of dispreferred markers may act as a subtle way to 'soften the blow' of bad news without the need to avoid delivering the information altogether.

(2) Consider using proportional quantifiers to communicate high severity risk to patients

It remains the case that terms like 'some' and 'many' are inherently ambiguous as they do not have a solid referent. As such, a speaker should always consider that when using a

proportional quantifier, the exact proportion that is represented by that quantifier is largely individual. However, a benefit of using these terms to qualify health risk, is that they do not appear to be unduly influenced by a severe outcome, resulting in our next proposition (2). In chapter five we looked at the interpretation of negative health statements quantified with 'a few', 'some', 'many' or 'most' (e.g., 'some people suffer…'), and discovered that the likelihood attached to experiencing both *insomnia* (low severity) and *deafness* (high severity) was the same for all quantifiers.

Considering that the severity bias (i.e., whereby more severe outcomes are considered more likely) affects the use of uncertainty terms (Bonnefon et al., 2006), as well as numerical quantifiers like the '1-in-X' ratio (see experiments reported in 5.3), our results suggest that the most effective way to combat the severity bias when communicating health risk is to do so by using proportional quantifiers. Perhaps this is one of the reasons why health professionals prefer using proportional quantifiers (Juanchich & Sirota, 2020a). A word of caution though, is that speakers using proportional quantifiers must consciously maintain an intention to be accurate (rather than trying to manage face). This is because the 'strength' of the quantifier chosen depends on what they intend to achieve with their speech (see experiment reported in 5.1), and altering explicitly communicated probability increases the potential for ambiguity and misunderstandings. Indeed, research has previously recommended that speakers make their intentions clear to their communicative partner, as a way to avoid mismatches of likelihood and misunderstandings (Sirota & Juanchich, 2012b).

One important caveat to the above propositions (1) and (2), is that they are not intended to be 'one size fits all'. Of course, patients have their own preferences on how they like to be communicated with (Parker et al., 2001), and in practice, the approach taken to deliver bad news should be personally tailored in all cases (Baile et al., 2000). Therefore,

statements (1) and (2) are better thought of as evidence-based suggestions, to be considered within context on an individual or situational basis (rather than as a universal).

7.6. Evaluation and Future Directions

Considering the eight experiments reported in this thesis, it is important to identify the factors that could limit the inferences we have drawn. The most prominent of these concerns is the use of hypothetical (imaginary) vignettes, rather than analysing genuine (spoken) language production and interpretation. One key objective of this thesis was to contribute to the current understanding of linguistic universals within a health context, so we must be cautious about making literal interpretations on the basis of 'presumed' pragmatic data (i.e., participants imagine the situation, rather than being physically in the situation).

The argument can be made then, that some of our findings do not adequately capture the way people *genuinely* communicate in the real world, because of the fictitious nature of our study designs. Rather than this however, we consider the opposite to be true for two reasons. First, the extent to which a person will act (or not act) through face-management concerns is stronger in-person than the same situation judged hypothetically. A study by Kawakami et al. (2009) analysed the difference between affective and behavioural responses to acts of racism. Across multiple studies it was found that participants would admit offence to a racist remark more frequently in a hypothetical situation than when faced with the real thing. Earlier data shows that the same is true for sexist comments, with aggrieved women more intent on confronting the one responsible in an hypothetical sense than was action shown through their responses to the real thing (Woodzicka & LaFrance, 2001). These data indicate that concerns with face (i.e., the decision not to express upset) are less influential when the situation is imaginary, and supports the argument that politeness-related effects are likely to be stronger in the real world (Sirota & Juanchich, 2015).

Second, the data we collected took place under experimental control, so language produced spontaneously is likely to exaggerate the effects we observed. Take, for example, the data we present in chapter four (part 1), where participants are tasked with writing a letter of diagnosis to a patient. We would be unable to conclude from this data that the language used in participants' letters are an exact literal reflection of their real-world language use. However, given that participants were able to re-visit, edit and proofread their letters before submitting, the finalised letters are actually a reflection of what they would have ideally wanted to say, if they were faced with that situation themselves (i.e., perfect, un-hesitant speech). In contrast, free speech is un-polished and imperfect, meaning the differences between our conditions are likely to be magnified under the spontaneity of real-world circumstances.

There are two study design elements that are important to consider when discussing the limiting factors of this thesis. The first is the implementation of a between-subjects (independent groups) design in the reported experiments. Allocating participants to only one of many possible conditions posed several challenges for addressing our research questions. Namely, independent groups designs introduce more variability (randomness) into a sample (Cook et al., 2002), and by extension, are less powerful than within-subjects designs (Fraley & Vazire, 2014). As discussed, many of the experiments reported in this thesis are highly powered, which on balance is likely to have alleviated (or 'evened out') the additional design-induced variability. We also believe that presenting participants with (in many cases) only one experimental vignette was the most effective way to both maximise engagement (outlined in section 3.5), and avoiding demand characteristics (i.e., masking the experimental manipulation).

The second design element that could limit our findings is the use of single-item measures. For example, in Experiment 8 (section 6.2) we measured the extent to which

participants considered the findings of a research study to be important. Our measure of importance was a single item (i.e., 'please give your best guess of how important this finding is'), and this is potentially problematic because a single-item measure is less reliable than a multi-item alternative (Nunnally, 1994). Consequently, our measure is more susceptible to measurement error (e.g., individual differences in what is considered to be 'important' research; Podsakoff et al., 2003). It is important to identify that our measures are not arbitrary, however. In line with the manifesto for slow science (outlined in section 3.6) we took steps to ensure that our methodology was reflective of established peer-reviewed work. For instance, our measure of importance in Experiment 8 was modelled closely on the work of DeJesus et al. (2019). Still, the use of single-item measures is a limitation that, if addressed in future research, would improve the reliability and validity of the reported findings.

It is also necessary to consider the factors that may limit the generality of findings presented in this thesis (Simons et al., 2017). Generally speaking, there are three key constraints on generality to be addressed. First, our inclusion criteria for all studies required that participants speak fluent English. Of course, this means that our findings can only be extrapolated as far as English-speaking populations. Second, in most of our experiments we recruited via opportunity sampling, which means that, like most psychological research, our sample is strongly female across-the-board, with 76.2% (n = 3159) of our entire sample (N = 4143) identifying this way. Also for this reason, a disproportionate number of participants are likely to be undergraduate psychology (or Health and Life Sciences Faculty) students. On the one hand this is likely to have improved the quality of our data. For example, many nursing students stated via email that they took an interest Experiment 1 (writing a letter of diagnosis) due to its' relevance to them, and this is beneficial because, anecdotally, the raw data from Experiment 1 signifies high engagement with the task (i.e., letters are often thorough and

written conscientiously). On the other hand, though, these respondents hold ‘a stake’ in the healthcare domain, so we cannot be certain that the letters provided are reflective of individual (natural) language production, rather than merely a simulation of how a health professional ‘should’ communicate. Finally, as a consequence of our sampling strategy it is likely that our findings fall victim to the biases of WEIRD (Western, Educated, Industrialised, Rich, Democratic) sampling (Henrich et al., 2010a; Henrich et al., 2010b). What follows is a more critical consideration of these constraints, and the steps necessary to overcome such methodological obstacles.

Future psycholinguistic research should continue to establish the extent to which politeness-based speech generalises across both languages and cultures. Despite the initial claim from Brown and Levinson (1987), that speech acts designed in the interest of face-work are universally applied strategies, studies in the field of cross-cultural pragmatics have shown that some polite utterances are considered on the basis of culture-specific preferences (e.g., requests; Barron, 2008, Byon, 2004; Reiter, 2002). Similarly, there are discrepancies in the strength and interpretation of polite requests across languages and between native or non-native speakers (Schauer, 2004; Woodfield, 2008). As such, it is important for research to understand the pragmatic mechanisms that either overlap with, or deviate from non-western cultures and languages. Thankfully, the number of studied languages within politeness is growing (Spencer-Oatey & Kádár, 2021), and by bridging this gap, the potential for misunderstandings is reduced (Haugh & Chang, 2015).

Another valuable solution to the aforementioned limitations, is to consider controlling for the variance associated with (potential) differences between men and women in a sample. Compared to other conceptualisations of polite language, politeness theory makes little mention of sex difference (Chalupnik et al., 2017), and this is perhaps why it is seldom

considered in contemporary research. A theory proposed by Lakoff (1975), which predates politeness theory, argues that men and women's talk is different, and typically done for different purposes. It has been evidenced in some early work, for example, that women use politeness expressions differently to men through an increased sensitivity to face-threat (Brown, 1990). However, these findings came from a highly culture-specific set of circumstances. A far stronger argument for the existence of sex-specific politeness is provided by Holmes (2013), who evidences that on the whole, women are more polite than men. Such evidence points to the need for future psycholinguistic research to statistically account for any potential variability in the production or interpretation of language, as a function of sex. There is, however, an important caveat to this, which is that the extent to which politeness may differ between men and women is very complicated (Mills, 2003; Nakane, 2008), and should always be considered within the context of a specific research question.

Finally, a relevant consideration is that in a number of experiments reported in this thesis, we developed our own experimental vignette that differed notably from those that have been developed previously. For example, the first experiment reported in chapter six places participants in the position of a news media personnel, and tasks them with choosing a headline for a health-related news article. Unlike other vignettes that are built closely on previously validated methods (e.g., chapter five), scenarios like this are 'self-engineered' to a greater extent. That is, rather than making small adjustments (e.g., adapting a vignette to a health context), we made some vignettes more so 'from scratch'. Where appropriate, steps were taken to ensure that our manipulations or measures were effective. For example, in chapter five we completed a manipulation check, confirming via a *t*-test that '*severe acne*' was considered to be significantly more serious than '*mild acne*' ($d = 1.27$). Similarly in chapter five, we confirmed that the proportion represented by our choice quantifiers arranged

themselves in the order that we expected (i.e., a few, some, many, most; Pezzelle et al., 2018a). Still, it remains true that some of our vignettes are less reliant on established peer-reviewed methods, so further study is needed to confirm the utility of the health-based scenarios we created.

As it applies to health communication, it would be useful for research to test whether the evidence-based recommendations made in section 7.5 have the potential to improve actual patient satisfaction. For instance, it appears based on the evidence reported in this thesis, that managing face-threat through indirectness is more effective than using explicit uncertainty terms (e.g., hedging), because it achieves face-work without unduly impacting the communicated probability. However, it is yet to be established whether being given indirect bad news leaves the (hypothetical) patient more satisfied than if they were given bad news through explicit hedging. Perhaps testing this directly would be valuable in understanding the potential for our recommendations to improve real-world patient experiences.

7.7. Overall Conclusions

The use of ambiguous language is problematic when communicating about human health because it causes a discrepancy between what a sentence *says* and what it actually *means*. This thesis set out to better understand the pragmatic (contextual) factors that influence both the production, and the comprehension of vague (or cryptic) language when exchanging health information.

On the production of health-based language, our findings show that when the situation is particularly threatening, a speaker who is communicating uncertainty will choose to blunt this threat by being subtly indirect, rather than by being explicitly uncertain (e.g., hedging). When instead using a quantifier to communicate threatening uncertainty (e.g., some or many),

a speaker will soften the impact of the information by expressing the event as less likely to happen. Finally, when a writer is communicating en masse (i.e., to an audience rather than an individual), they will decide to use a generic statement far more often when they are motivated to communicate in an appealing way.

On the comprehension of health-based language, we show that indirectly delivered threatening news does not alter the communicated probability of the event occurring, and does not alter the perceived certainty of the speaker. When uncertainty is communicated with a verbal quantifier (e.g., a few or most), severe events are not judged as more likely to occur. This is in-contrast with '1-in-X' ratios (e.g., 1 in 10 chance of suffering side effects), which cause a person to overestimate the risk of a severe event occurring. Lastly, when reading about health research in the media, the use of a universal statement to summarise this causes an overstatement of the importance and generality of research. Our overall findings serve as a reminder that the contextual elements of a social situation largely determine the language used and understood within it.

This thesis applied the principles of experimental pragmatics to further the current understanding of psycholinguistic theory. This was achieved by manipulating health-based situations, which provided the ideal test-bed for inducing the effects of politeness and face-work that we aimed to discover. We also extend previous works in terms of applicability, because exchanging information about health is relevant to everyone. Future work should seek to replicate and extend the findings presented in this thesis, in the hope of achieving a more fine-grained understanding of the psychological mechanisms that underpin human interaction.

References

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., & Gernsbacher, M. A. (2020). A consensus-based transparency checklist. *Nature human behaviour*, *4*(1), 4-6. https://doi.org/10.1038/s41562-019-0772-6

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational research methods*, *17*(4), 351-371. https://doi.org/10.1177/1094428114547952

Ahern, N. R. (2005). Using the Internet to conduct research. *Nurse researcher*, *13*(2).

Aklin, M., & Urpelainen, J. (2014). Perceptions of scientific dissent undermine public support for environmental policy. *Environmental Science & Policy*, *38*, 173-177. https://doi.org/10.1016/j.envsci.2013.10.006

Almond, S., Mant, D., & Thompson, M. (2009). Diagnostic safety-netting. *British Journal of General Practice*, *59*(568), 872-874. https://doi.org/10.3399/bjgp09X472971

Aronsson, K., & Sätterlund-Larsson, U. (1987). Politeness strategies and doctor-patient communication. On the social choreography of collaborative thinking. *Journal Of Language and Social Psychology*, *6*(1), 1-27.

Arthur Jr, W., Hagen, E., & George Jr, F. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*, 105-137.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(3), 128.

Auger, A., & Roy, J. (2008). Expression of uncertainty in linguistic data. In *2008 11th International Conference on Information Fusion* (pp. 1-8). IEEE.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior research methods*, *45*(2), 527-535. https://doi.org/10.3758/s13428-012-0265-2

Baile, W. F., & Beale, E. A. (2003). Giving bad news to cancer patients: matching process and content. *Journal of clinical oncology*, *21*(90090), 49-51. https://doi.org/10.1200/JCO.2003.01.169

Baile, W. F., Buckman, R., Lenzi, R., Glober, G., Beale, E. A., & Kudelka, A. P. (2000). SPIKES—a six-step protocol for delivering bad news: application to the patient with cancer. In (Vol. 5, pp. 302-311): Oxford University Press.

Baker, M. (2016a). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452-454. https://doi.org/10.1038/533452a

Baker, M. (2016b). Reproducibility crisis. *Nature*, *533*(26), 353-366. https://doi.org/10.1038/533452a

Barclay, J. S., Blackhall, L. J., & Tulsky, J. A. (2007). Communication strategies and cultural issues in the delivery of bad news. *Journal of palliative medicine*, *10*(4), 958-977. https://doi.org/10.1089/jpm.2007.9929

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Barron, A. (2008). Contrasting requests in Inner Circle Englishes: A study in variational pragmatics. *Developing contrastive pragmatics: Interlanguage and cross-cultural perspectives*, 355-402.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of general psychology*, *5*(4), 323-370.

Baxter, L. A. (1984). An investigation of compliance-gaining as politeness. *Human communication research*, *10*(3), 427-456. https://doi.org/DOI 10.1111/j.1468-2958.1984.tb00026.x

Beck, S., Bergenholtz, C., Bogers, M., Brasseur, T.-M., Conradsen, M. L., Di Marco, D., Distel, A. P., Dobusch, L., Dörler, D., & Effert, A. (2022). The Open Innovation in Science research field: a collaborative conceptualisation approach. *Industry and Innovation*, *29*(2), 136-185.

Beck, S., Mahdad, M., Beukel, K., & Poetz, M. (2019). The value of scientific knowledge dissemination for scientists—A value capture perspective. *Publications*, *7*(3), 54. https://doi.org/ARTN 5410.3390/publications7030054

Berkey, F. J., Wiedemer, J. P., & Vithalani, N. D. (2018). Delivering bad or life-altering news. *American family physician*, *98*(2), 99-104. https://www.ncbi.nlm.nih.gov/pubmed/30215989

Beukeboom, C. J., Finkenauer, C., & Wigboldus, D. H. (2010). The negation bias: When negations signal stereotypic expectancies. *Journal of personality and social psychology*, *99*(6), 978. https://doi.org/10.1037/a0020861

Bhise, V., Meyer, A. N., Menon, S., Singhal, G., Street, R. L., Giardina, T. D., & Singh, H. (2018). Patient perspectives on how physicians communicate diagnostic uncertainty: an experimental vignette study. *International Journal for Quality in Health Care*, *30*(1), 2-8. https://doi.org/10.1093/intqhc/mzx170

Blackhall, L. J., Frank, G., Murphy, S., & Michel, V. (2001). Bioethics in a different tongue: the case of truth-telling. *Journal of Urban Health*, *78*(1), 59-71. https://doi.org/10.1093/jurban/78.1.59

Blanch, D. C., Hall, J. A., Roter, D. L., & Frankel, R. M. (2009). Is it good to express uncertainty to a patient? Correlates and consequences for medical students in a standardized patient visit. *Patient education and counseling*, *76*(3), 300-306.

Blum-Kulka, S., Danet, B., & Gherson, R. (1985). The language of requesting in Israeli society. In *Language and social situations* (pp. 113-139). Springer.

Bonett, D. G. (2012). Replication-extension studies. *Current Directions in Psychological Science*, *21*(6), 409-412. https://doi.org/10.1177/0963721412459512

Bonnefon, J.-F., Dahl, E., & Holtgraves, T. M. (2015). Some but not all dispreferred turn markers help to interpret scalar terms in polite contexts. *Thinking & Reasoning*, *21*(2), 230-249.

Bonnefon, J.-F., De Neys, W., & Feeney, A. (2011a). Processing scalar inferences in face-threatening contexts. *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33)

Bonnefon, J.-F., Feeney, A., & De Neys, W. (2011b). The risk of polite misunderstandings. *Current Directions in Psychological Science*, *20*(5), 321-324. https://doi.org/10.1177/0963721411418472

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249-258. https://doi.org/10.1016/j.cognition.2009.05.005

Bonnefon, J.-F., & Villejoubert, G. (2006). Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychological Science*, *17*(9), 747-751. https://doi.org/10.1111/j.1467-9280.2006.01776.x

Bossema, F. G., Burger, P., Bratton, L., Challenger, A., Adams, R. C., Sumner, P., Schat, J., Numans, M. E., & Smeets, I. (2019). Expert quotes and exaggeration in health news: a retrospective quantitative content analysis. *Wellcome Open Research*, *4*, 56. https://doi.org/10.12688/wellcomeopenres.15147.2

Brandner, J., Behne, M., Huesing, B., & Moll, I. (2006). Caffeine improves barrier function in male skin. *International journal of cosmetic science*, *28*(5), 343-347. https://doi.org/10.1111/j.1467-2494.2006.00346.x

Bratton, L., Adams, R. C., Challenger, A., Boivin, J., Bott, L., Chambers, C. D., & Sumner, P. (2020). Causal overstatements reduced in press releases following academic study of health news. *Wellcome Open Research*, *5*, 6. https://doi.org/10.12688/wellcomeopenres.15647.2

Bromme, R., Brummernhenrich, B., Becker, B. M., & Jucks, R. (2012). The effects of politeness-related instruction on medical tutoring. *Communication Education*, *61*(4), 358-379.

Brown, P. (1990). Gender, politeness, and confrontation in Tenejapa. *Discourse Processes*, *13*(1), 123-141. https://doi.org/Doi 10.1080/01638539009544749

Brown, P., & Levinson, S. C. (1978). Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction* (pp. 56-311). Cambridge University Press.

Brown, P., Levinson, S. C., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Brown, R., & Gilman, A. (1989). Politeness theory and Shakespeare's four major tragedies. *Language in society*, *18*(2), 159-212.

Brown, V., Parker, P., Furber, L., & Thomas, A. (2011). Patient preferences for the delivery of bad news–the experience of a UK Cancer Centre. *European journal of cancer care*, *20*(1), 56-61. https://doi.org/10.1111/j.1365-2354.2009.01156.x

Brummernhenrich, B., & Jucks, R. (2019). "Get the shot, now!" Disentangling content-related and social cues in physician–patient communication. *Health Psychology Open*, *6*(1), 2055102919833057.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of cognition*.

Buckman, R. A. (2005). Breaking bad news: the SPIKES strategy. *Community oncology*, *2*(2), 138-142.

Buhse, S., Rahn, A. C., Bock, M., & Mühlhauser, I. (2018). Causal interpretation of correlational studies–Analysis of medical news on the website of the official journal for German physicians. *PloS one*, *13*(5), e0196833. https://doi.org/10.1371/journal.pone.0196833

Burgers, C., Beukeboom, C. J., & Sparks, L. (2012). How the doc should (not) talk: When breaking bad news with negations influences patients' immediate responses and medical adherence intentions. *Patient education and counseling*, *89*(2), 267-273.

Byon, A. S. (2004). Sociopragmatic analysis of Korean requests: Pedagogical settings. *Journal of Pragmatics*, *36*(9), 1673-1704. https://doi.org/10.1016/j.pragma.2004.05.003

Chalupnik, M., Christie, C., & Mullany, L. (2017). (Im) politeness and gender. In *The Palgrave handbook of linguistic (im) politeness* (pp. 517-537). Springer.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., ... & De Rosario, M. H. (2020). Package 'pwr'. *R package version*, *1*(2).

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). Evaluation of negation phrases in narrative clinical reports. Proceedings of the AMIA Symposium,

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, *34*(8), 1452-1482. https://doi.org/10.1111/j.1551-6709.2010.01126.x

Clelland, H. T. & Haigh, M. (2023). Materials and Data from Politeness and the Communication of Uncertainty when Breaking Bad News [Data set & Materials]. Open Science Framework. https://doi.org/10.17605/OSF.IO/ZU2AN

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (pp. 103-134). Boston, MA: Houghton Mifflin.

Coste, J., Quinquis, L., Audureau, E., & Pouchot, J. (2013). Non response, incomplete and inconsistent responses to self-administered health-related quality of life measures in the general population: patterns, determinants and impact on the validity of estimates—a population-based study in France using the MOS SF-36. *Health and quality of life outcomes*, *11*(1), 1-15. https://doi.org/Artn 4410.1186/1477-7525-11-44

Cousin, G., Schmid Mast, M., & Jaunin-Stalder, N. (2013). When physician-expressed uncertainty leads to patient dissatisfaction: a gender study. *Medical education*, *47*(9), 923-931.

Crompton, P. (1997). Hedging in academic writing: Some theoretical problems. *English for specific purposes*, *16*(4), 271-287.

Cushing, S. (1994). *Fatal words: Communication clashes and aircraft crashes*. University of Chicago Press.

Czarnitzki, D., Grimpe, C., & Pellens, M. (2015). Access to research inputs: open science versus the entrepreneurial university. *The Journal of Technology Transfer*, *40*(6), 1050-1063. https://doi.org/10.1007/s10961-015-9392-0

Dalianis, H., & Velupillai, S. (2010). How certain are clinical assessments? Annotating Swedish clinical text for (un) certainties, speculations and negations. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10),

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraint-based approach. *Cognitive science*, *39*(4), 667-710. https://doi.org/10.1111/cogs.12171

DeJesus, J. M., Callanan, M. A., Solis, G., & Gelman, S. A. (2019). Generic language in scientific communication. *Proceedings of the National Academy of sciences*, *116*(37), 18370-18377. https://doi.org/10.1073/pnas.1817706116

Del Vento, A., Bavelas, J., Healing, S., MacLean, G., & Kirk, P. (2009). An experimental investigation of the dilemma of delivering bad news. *Patient education and counseling*, *77*(3), 443-449. https://doi.org/10.1016/j.pec.2009.09.014

Dibble, J. L., & Levine, T. R. (2013). Sharing good and bad news with friends and strangers: Reasons for and communication behaviors associated with the MUM effect. *Communication Studies*, *64*(4), 431-452.

Duda, M. D., & Nobile, J. L. (2010). The fallacy of online surveys: No data are better than bad data. *Human Dimensions of Wildlife*, *15*(1), 55-64.

Dumas-Mallet, E., Smith, A., Boraud, T., & Gonon, F. (2018). Scientific uncertainty in the press: How newspapers describe initial biomedical findings. *Science communication*, *40*(1), 124-141. https://doi.org/10.1177/1075547017752166

Ellis, D. A. (2020). *Smartphones within psychological science*. Cambridge University Press.

Fagerlin, A., Zikmund-Fisher, B. J., & Ubel, P. A. (2011). Helping patients decide: ten steps to better risk communication. *Journal of the National Cancer Institute*, *103*(19), 1436-1443. https://doi.org/10.1093/jnci/djr318

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

Feeney, A., & Bonnefon, J.-F. (2013). Politeness and honesty contribute additively to the interpretation of scalar expressions. *Journal of language and social psychology*, *32*(2), 181-190. https://doi.org/10.1177/0261927x12456840

Fischer, K., & Jungermann, H. (1996). Rarely occurring headaches and rarely occurring blindness: Is rarely= rarely? The meaning of verbal frequentistic labels in specific medical contexts. *Journal of Behavioral Decision Making*, *9*(3), 153-172.

Forgas, J. P. (1999a). Feeling and speaking: Mood effects on verbal communication strategies. *Personality and Social Psychology Bulletin*, *25*(7), 850-863.

Forgas, J. P. (1999b). On feeling good and being rude: Affective influences on language use and request formulations. *Journal of personality and social psychology*, *76*(6), 928. https://doi.org/Doi 10.1037/0022-3514.76.6.928

Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association: JMLA*, *105*(2), 203.

Fraenkel, T., & Schul, Y. (2008). The meaning of negated adjectives. *Intercultural Pragmatics*, *5*(4), 517-540. https://doi.org/10.1515/Iprg.2008.025

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Fraser, B. (2010). Pragmatic competence: The case of hedging. *New approaches to hedging*, *1534*.

Frith, U. (2020). Fast lane to slow science. *Trends in cognitive sciences*, *24*(1), 1-2. https://doi.org/10.1016/j.tics.2019.10.007

Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, *22*(2), 313.

Garnham, A. (2018). Pragmatics and Inference.

Gelman, S. A., & Roberts, S. O. (2017). How language shapes the cultural inheritance of categories. *Proceedings of the National Academy of sciences*, *114*(30), 7900-7907. https://doi.org/10.1073/pnas.1621073114

Gibson, A. M., & Bowling, N. A. (2020). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, *36*(2), 410. https://doi.org/10.1027/1015-5759/a000526

Giora, R., Balaban, N., Fein, O., & Alkabets, I. (2004). Explicit negation as positivity in disguise. In *Figurative language comprehension* (pp. 245-270). Routledge.

Girgis, A., & Sanson-Fisher, R. W. (1995). Breaking bad news: consensus guidelines for medical practitioners. *Journal of clinical oncology*, *13*(9), 2449-2456. https://doi.org/10.1200/JCO.1995.13.9.2449

Gladwell, M. (2008). *Outliers: The story of success*. Little, Brown.

Goffman, E. (1967). Interaction ritual: Essays on face-to-face interaction.

Goffman, E. (1971). Normal Appearances. Relations in Public. E. Goffman. In: New York, Harper and Row.

Gonzales, M. H., Pederson, J. H., Manning, D. J., & Wetter, D. W. (1990). Pardon my gaffe: Effects of sex, status, and consequence severity on accounts. *Journal of personality and social psychology*, *58*(4), 610. https://doi.org/Doi 10.1037/0022-3514.58.4.610

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, *8*(341), 341ps312-341ps312.

Gordon, G. H., Joos, S. K., & Byrne, J. (2000). Physician expressions of uncertainty during patient encounters. *Patient education and counseling*, *40*(1), 59-65. https://doi.org/10.1016/s0738-3991(99)00069-5

Granello, D. H., & Wheaton, J. E. (2004). Online data collection: Strategies for research. *Journal of Counseling & Development*, *82*(4), 387-393. https://doi.org/DOI 10.1002/j.1556-6678.2004.tb00325.x

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.

Gutiérrez, K. D., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. *Educational researcher*, *32*(5), 19-25.

Haigh, M., Birch, H. A., & Clelland, H. T. (2022). “Experts agree…” The production and comprehension of consensus generics. Manuscript in preparation.

Haigh, M., Birch, H. A., & Pollet, T. V. (2020). Does 'Scientists Believe…'Imply 'All Scientists believe…'? Individual Differences in the Interpretation of Generic News Headlines. *Collabra: Psychology*, *6*(1).

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23-34.

Hamilton, R., Vohs, K. D., & McGill, A. L. (2014). We'll be honest, this won't be the best article you'll ever read: The use of dispreferred markers in word-of-mouth communication. *Journal of Consumer Research*, *41*(1), 197-212. https://doi.org/10.1086/675926

Harris, A. J., & Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1571. https://doi.org/10.1037/a0024195

Haugh, M., & Chang, W.-L. M. (2015). Understanding im/politeness across cultures: an interactional approach to raising sociopragmatic awareness. *International Review of Applied Linguistics in Language Teaching*, *53*(4), 389-414. https://doi.org/10.1515/iral-2015-0018

Helft, P. R., & Petronio, S. (2007). Communication pitfalls with cancer patients:"hit-and-run" deliveries of bad news. *Journal of the American College of Surgeons*, *205*(6), 807-811. https://doi.org/10.1016/j.jamcollsurg.2007.07.022

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature*, *466*(7302), 29-29.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? *Behavioral and brain sciences*, *33*(2-3), 61-83. https://doi.org/10.1017/S0140525X0999152X

Henry, M. S. (2006). Uncertainty, responsibility, and the evolution of the physician/patient relationship. *Journal of Medical Ethics*, *32*(6), 321-323. https://doi.org/10.1136/jme.2005.013987

Hoerger, M. (2010). Participant dropout as a function of survey length in Internet-mediated university studies: Implications for study design and voluntary participation in psychological research. *Cyberpsychology, behavior, and social networking*, *13*(6), 697-700.

Holmes, J. (2013). *Women, men and politeness*. Routledge.

Holtgraves, T. (2005). Social psychology, cognitive psychology, and linguistic politeness.

Holtgraves, T. (2014). Interpreting uncertainty terms. *Journal of personality and social psychology*, *107*(2), 219. https://doi.org/10.1037/a0036930

Holtgraves, T., & Bonnefon, J.-F. (2017). Experimental approaches to linguistic (im) politeness. In *The Palgrave Handbook of Linguistic (Im) politeness* (pp. 381-401). Springer.

Holtgraves, T., & Joong-Nam, Y. (1990). Politeness as universal: Cross-cultural perceptions of request strategies and inferences based on their use. *Journal of personality and social psychology*, *59*(4), 719. https://doi.org/Doi 10.1037/0022-3514.59.4.719

Holtgraves, T., & Perdew, A. (2016). Politeness and the communication of uncertainty. *Cognition*, *154*, 1-10. https://doi.org/10.1016/j.cognition.2016.05.005

Holtgraves, T., & Yang, J.-N. (1992). Interpersonal underpinnings of request strategies: General principles and differences due to culture and gender. *Journal of personality and social psychology*, *62*(2), 246. https://doi.org/Doi 10.1037/0022-3514.62.2.246

Holtgraves, T., & Yang, J.-N. (1992). Interpersonal underpinnings of request strategies: General principles and differences due to culture and gender. *Journal of personality and social psychology*, *62*(2), 246. https://doi.org/Doi 10.1037/0022-3514.62.2.246

Horn, L. (1984). Towards a new taxonomy for pragmatic inference: Q-and R-based implicature. *Meaning, form and use in context*.

Horn, L. R. (1992). The said and the unsaid. Semantics and Linguistic Theory,

Hossain, M. A., Dwivedi, Y. K., & Rana, N. P. (2016). State-of-the-art in open data research: Insights from existing literature and a research agenda. *Journal of organizational computing and electronic commerce*, *26*(1-2), 14-40. https://doi.org/Doi 10.1080/10919392.2015.1124007

Ioannidis, J. P. (2018). Why replication has more scientific value than original discovery. *Behavioral and brain sciences*, *41*. https://doi.org/ARTN e13710.1017/S0140525X18000729

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524-532. https://doi.org/10.1177/0956797611430953

Johnson, C. G., Levenkron, J. C., Suchman, A. L., & Manchester, R. (1988). Does physician uncertainty affect patient satisfaction? *Journal of general internal medicine*, *3*(2), 144-149.

Juanchich, M., & Sirota, M. (2013). Do people really say it is “likely” when they believe it is only “possible”? Effect of politeness on risk communication. In: SAGE Publications Sage UK: London, England.

Juanchich, M., & Sirota, M. (2020a). Most family physicians report communicating the risks of adverse drug reactions in words (vs. numbers). *Applied Cognitive Psychology*, *34*(2), 526-534. https://doi.org/10.1002/acp.3623

Juanchich, M., & Sirota, M. (2020b). Do people really prefer verbal probabilities? *Psychological research*, *84*(8), 2325-2338.

Juanchich, M., Sirota, M., & Bonnefon, J.-F. (2019). Verbal uncertainty.

Juanchich, M., Sirota, M., & Butler, C. L. (2012). The perceived functions of linguistic risk quantifiers and their effect on risk, negativity perception and decision making. *Organizational Behavior and Human Decision Processes*, *118*(1), 72-81. https://doi.org/10.1016/j.obhdp.2012.01.002

Kakai, H. (2002). A double standard in bioethical reasoning for disclosure of advanced cancer diagnoses in Japan. *Health Communication*, *14*(3), 361-376. https://doi.org/10.1207/S15327027HC1403_4

Katsos, N. (2008). The semantics/pragmatics interface from an experimental perspective: the case of scalar implicature. *Synthese*, *165*(3), 385-401. https://doi.org/10.1007/s11229-007-9187-4

Kawakami, K., Dunn, E., Karmali, F., & Dovidio, J. F. (2009). Mispredicting affective and behavioral responses to racism. *Science*, *323*(5911), 276-278. https://doi.org/10.1126/science.1164951

Keenan, E. L., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and philosophy*, *9*(3), 253-326. https://doi.org/Doi 10.1007/Bf00630273

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., & Bahník, Š. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490. https://doi.org/10.1177/2515245918810225

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., & Bahník, Š. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and*

Practices in Psychological Science, *1*(4), 443-490. https://doi.org/10.1177/2515245918810225

Klovning, A., Sandvik, H., & Hunskaar, S. (2009). Web-based survey attracted age-biased sample with more severe illness than paper-based survey. *Journal of clinical epidemiology*, *62*(10), 1068-1074. https://doi.org/10.1016/j.jclinepi.2008.10.015

Krifka, M., Pelletier, F. J., Carlson, G., Ter Meulen, A., Chierchia, G., & Link, G. (1995). Genericity: an introduction.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*, 1-26.

Lambert, B. L. (1996). Face and politeness in pharmacist-physician interaction. *Social science & medicine*, *43*(8), 1189-1198. https://doi.org/10.1016/0277-9536(95)00370-3

Lazaridou-Chatzigoga, D. (2019). Genericity.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., & Rothschild, D. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

Ledford, H. (2015). Team science. *Nature*, *525*(7569), 308. https://doi.org/DOI10.1038/525308a

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: the theory of the strategic speaker. *Psychological review*, *117*(3), 785. https://doi.org/10.1037/a0019688

Lefever, S., Dal, M., & Matthíasdóttir, Á. (2007). Online data collection in academic research: advantages and limitations. *British Journal of Educational Technology*, *38*(4), 574-582. https://doi.org/10.1111/j.1467-8535.2006.00638.x

Leichty, G., & Applegate, J. L. (1991). Social-cognitive and situational influences on the use of face-saving persuasive strategies. *Human communication research*, *17*(3), 451-484.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2022). Emmeans: Estimated marginal means, aka least-squares means. *R package version*, *1*(1), 3.

Leslie, S.-J. (2007). Generics and the structure of the mind. *Philosophical perspectives*, *21*(1), 375-403. https://doi.org/10.1111/j.1520-8583.2007.00138.x

Leslie, S.-J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1), 1-47. https://doi.org/10.1215/00318108-2007-023

Leslie, S.-J., Khemlani, S., & Glucksberg, S. (2011). Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language*, *65*(1), 15-31.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Levinson, W., Stiles, W. B., Inui, T. S., & Engle, R. (1993). Physician frustration in communicating with patients. *MEDICAL CARE-PHILADELPHIA-*, *31*(4), 285-285. https://doi.org/10.1097/00005650-199304000-00001

Ley, P. (1982). Giving information to patients. *Social Psychology and Bhavioural Science*.

Lim, T. S., & Bowers, J. W. (1991). Facework solidarity, approbation, and tact. *Human communication research*, *17*(3), 415-450. https://doi.org/DOI 10.1111/j.1468-2958.1991.tb00239.x

Liu, D., Juanchich, M., Sirota, M., & Orbell, S. (2020). The intuitive use of contextual information in decisions made with verbal and numerical quantifiers. *Quarterly Journal of Experimental Psychology*, *73*(4), 481-494. https://doi.org/10.1177/1747021820903439

Liu, D., Juanchich, M., Sirota, M., & Orbell, S. (2021). Differences between decisions made using verbal or numerical quantifiers. *Thinking & Reasoning*, *27*(1), 69-96.

Loy, J. E., Rohde, H., & Corley, M. (2019). Real-time social reasoning: the effect of disfluency on the meaning of some. *Journal of Cultural Cognitive Science*, *3*(2), 159-173. https://doi.org/10.1007/s41809-019-00037-1

Mast, M. S., Kindlimann, A., & Langewitz, W. (2005). Recipients' perspective on breaking bad news: How you put it really makes a difference. *Patient education and counseling*, *58*(3), 244-251. https://doi.org/10.1016/j.pec.2005.05.005

Mazzarella, D., Trouche, E., Mercier, H., & Noveck, I. (2018). Believing what you're told: Politeness and scalar inferences. *Frontiers in psychology*, *9*, 908. https://doi.org/ARTN 90810.3389/fpsyg.2018.00908

Mclaughun, M. L., Cody, M. J., & O'Hair, H. D. (1983). The management of failure events: Some contextual determinants of accounting behavior. *Human communication research*, *9*(3), 208-224.

Miceli, S. (2019). Reproducibility and replicability in research. In.

Mills, S. (2003). *Gender and politeness*. Cambridge University Press.

Miyaji, N. T. (1993). The power of compassion: truth-telling among American doctors in the care of dying patients. *Social science & medicine*, *36*(3), 249-264. https://doi.org/Doi10.1016/0277-9536(93)90008-R

Moore, D. A. (2016). Preregister if you want to. *American Psychologist*, *71*(3), 238. https://doi.org/10.1037/a0040195

Moxey, L. M., & Sanford, A. J. (1986). Quantifiers and focus. *Journal of semantics*, *5*(3), 189-206.

Moxey, L. M., & Sanford, A. J. (2000). Communicating quantities: A review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *14*(3), 237-255. https://doi.org/Doi10.1002/(Sici)1099-0720(200005/06)14:3<237::Aid-Acp641>3.0.Co;2-R

Musch, J., & Klauer, K. C. (2002). Psychological experimenting on the World Wide Web: Investigating content effects in syllogistic reasoning. *Online social sciences*, *1*, 181-212.

Nakane, I. (2008). Politeness and gender in interpreted police interviews. *Monash University linguistics papers*, *6*(1), 29-40.

NHS (2019, August 13). Caesarean section. NHS UK. https://www.nhs.uk/conditions/caesarean-section/

Nouwen, R. (2010). What's in a quantifier? *The Linguistics Enterprise: From knowledge of language to knowledge in linguistics*, *150*, 235.

Noveck, I. (2018). *Experimental pragmatics: The making of a cognitive science*. Cambridge University Press.

Noveck, I., & Sperber, D. (2012). 14 The why and how of experimental pragmatics: the case of 'scalar inferences'. *Meaning and relevance*, 307.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165-188. https://doi.org/Doi10.1016/S0010-0277(00)00114-1

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and language*, *85*(2), 203-210. https://doi.org/10.1016/s0093-934x(03)00053-1

Nunnally, J. C. (1994). The assessment of reliability. *Psychometric theory*.

Okamoto, S., & Robinson, W. P. (1997). Determinants of gratitude expressions in England. *Journal of language and social psychology*, *16*(4), 411-433. https://doi.org/Doi 10.1177/0261927x970164003

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, *45*(4), 867-872. https://doi.org/10.1016/j.jesp.2009.03.009

P Simmons, J., D Nelson, L., & Simonsohn, U. (2021). Pre-registration: Why and how. *Journal of Consumer Psychology*, *31*(1), 151-162.

Pain, E. (2016). How to (seriously) read a scientific paper. *Science, 10*.

Pan, Y., & Zhan, P. (2020). The impact of sample attrition on longitudinal learning diagnosis: A prolog. *Frontiers in psychology*, *11*, 1051. https://doi.org/10.3389/fpsyg.2020.01051

Parker, P. A., Baile, W. F., de Moor, C., Lenzi, R., Kudelka, A. P., & Cohen, L. (2001). Breaking bad news about cancer: patients' preferences for communication. *Journal of clinical oncology*, *19*(7), 2049-2056. https://doi.org/Doi 10.1200/Jco.2001.19.7.2049

Paul, C., Clinton-McHarg, T., Sanson-Fisher, R., Douglas, H., & Webb, G. (2009). Are we there yet? The state of the evidence base for guidelines on breaking bad news to cancer patients. *European Journal of Cancer*, *45*(17), 2960-2966.

Penn, C., Watermeyer, J., & Evans, M. (2011). Why don't patients take their drugs? The role of communication, context and culture in patient adherence and the work of the pharmacist in HIV/AIDS. *Patient education and counseling*, *83*(3), 310-318.

Person, N. K., Kreuz, R. J., Zwaan, R. A., & Graesser, A. C. (1995). Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. *Cognition and Instruction*, *13*(2), 161-188.

Peters, U., Krauss, A., & Braganza, O. (2022). Generalization bias in science. *Cognitive science*, *46*(9), e13188. https://doi.org/10.1111/cogs.13188

Pezzelle, S., Bernardi, R., & Piazza, M. (2018a). Probing the mental representation of quantifiers. *Cognition*, *181*, 117-126. https://doi.org/10.1016/j.cognition.2018.08.009

Pezzelle, S., Steinert-Threlkeld, S., Bernardi, R., & Szymanik, J. (2018b). Some of them can be guessed! Exploring the effect of linguistic context in predicting quantifiers. *arXiv preprint arXiv:1806.00354*, 114-119. <Go to ISI>://WOS:000493913100019

Pighin, S., & Bonnefon, J.-F. (2011). Facework and uncertain reasoning in health communication. *Patient education and counseling*, *85*(2), 169-172. https://doi.org/10.1016/j.pec.2010.09.005

Pighin, S., Savadori, L., Barilli, E., Cremonesi, L., Ferrari, M., & Bonnefon, J.-F. (2011). The 1-in-X effect on the subjective assessment of medical probabilities. *Medical Decision Making*, *31*(5), 721-729. https://doi.org/10.1177/0272989X11403490

Pighin, S., Savadori, L., Barilli, E., Galbiati, S., Smid, M., Ferrari, M., & Cremonesi, L. (2015). Communicating Down syndrome risk according to maternal age:"1-in-X" effect on perceived risk. *Prenatal diagnosis*, *35*(8), 777-782. https://doi.org/10.1002/pd.4606

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of sciences*, *105*(3), 833-838. https://doi.org/10.1073/pnas.0707192105

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, *88*(5), 879.

Politi, M. C., Han, P. K., & Col, N. F. (2007). Communicating the uncertainty of harms and benefits of medical interventions. *Medical Decision Making*, *27*(5), 681-695. https://doi.org/10.1177/0272989X07307270

Ptacek, J. T., & Eberhardt, T. L. (1996). Breaking bad news: a review of the literature. *Jama*, *276*(6), 496-502. https://www.ncbi.nlm.nih.gov/pubmed/8691562

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. *References-Scientific Research Publishing*. URL https://www.R-project.org/.

Quill, T. E., & Townsend, P. (1991). Bad news: delivery, dialogue, and dilemmas. *Archives of internal medicine*, *151*(3), 463-468. https://www.ncbi.nlm.nih.gov/pubmed/2001128

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental psychology*, *49*(4), 243. https://doi.org/10.1026/1618-3169.49.4.243

Reips, U.-D. (2009). Internet experiments: Methods, guidelines, metadata. *Human vision and electronic imaging XIV* (Vol.7240, pp. 53-61). SPIE.

Reips, U.-D. (2011). Privacy and the Disclosure of Information on the Internet: Issues and Measurement. *Internet in psychological research*, 67-100.

Reiter, R. M. (2002). A contrastive study of conventional indirectness in Spanish: Evidence from Peninsular and Uruguayan Spanish. *Pragmatics*, *12*(2), 135-151.

Rodriguez, K. L., Gambino, F. J., Butow, P., Hagerty, R., & Arnold, R. M. (2007). Pushing up daisies: implicit and explicit language in oncologist–patient communication about

death. *Supportive care in cancer*, *15*(2), 153-161. https://doi.org/10.1007/s00520-006-0108-8

Rogoff, B. (2003). *The cultural nature of human development*. Oxford university press.

Rosen, S., & Tesser, A. (1970). On reluctance to communicate undesirable information: The MUM effect. *Sociometry*, 253-263.

RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management journal*, *43*(6), 1248-1264. https://doi.org/Doi 10.2307/1556348

Schat, J., Bossema, F. G., Numans, M. E., Smeets, I., & Burger, P. (2018). Exaggerated health news: association between exaggeration in university press releases and exaggeration in news media coverage. *Nederlands Tijdschrift voor Geneeskunde*, *162*(1).

Schauer, G. A. (2004). May you speak louder maybe? *EUROSLA yearbook*, *4*.

Schleyer, T. K., & Forrest, J. L. (2000). Methods for the design and administration of web-based surveys. *Journal of the American Medical Informatics Association*, *7*(4), 416-425. https://doi.org/10.1136/jamia.2000.0070416

Schul, Y. (2011). Alive or not dead: Implications for framing from research on negations. *Perspectives on framing*, *157*, 176.

Sennet, A. (2011). Ambiguity.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123-1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, *143*(2), 534. https://doi.org/10.1037/a0033242

Sirota, M., & Juanchich, M. (2012a). Risk communication on shaky ground. *Science*, *338*(6112), 1286-1287. https://doi.org/10.1126/science.338.6112.1286

Sirota, M., & Juanchich, M. (2012b). To what extent do politeness expectations shape risk perception? Even numerical probabilities are under their spell! *Acta Psychologica*, *141*(3), 391-399.

Sirota, M., & Juanchich, M. (2015). A direct and comprehensive test of two postulates of politeness theory applied to uncertainty communication. *Judgment and Decision Making*, *10*(3), 232-240. <Go to ISI>://WOS:000355331200004

Sirota, M., & Juanchich, M. (2019). Ratio format shapes health decisions: The practical significance of the "1-in-X" effect. *Medical Decision Making*, *39*(1), 32-40. https://doi.org/10.1177/0272989X18814256

Sirota, M., Juanchich, M., & Bonnefon, J.-F. (2018a). "1-in-X" bias:"1-in-X" format causes overestimation of health-related risks. *Journal of Experimental Psychology: Applied*, *24*(4), 431. https://doi.org/10.1037/xap0000190

Sirota, M., Juanchich, M., Kostopoulou, O., & Hanak, R. (2014). Decisive evidence on a smaller-than-you-think phenomenon: revisiting the "1-in-X" effect on subjective medical probabilities. *Medical Decision Making*, *34*(4), 419-429. https://doi.org/10.1177/0272989X13514776

Sirota, M., Juanchich, M., Petrova, D., Garcia-Retamero, R., Walasek, L., & Bhatia, S. (2018b). Health professionals prefer to communicate risk-related numerical information using "1-in-X" ratios. *Medical Decision Making*, *38*(3), 366-376.

Slugoski, B. R., & Turnbull, W. (1988). Cruel to be kind and kind to be cruel: Sarcasm, banter and social relations. *Journal of language and social psychology*, *7*(2), 101-121.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society open science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Snow, J., & Mann, M. (2013). Qualtrics survey software: handbook for research professionals. *Qualtrics Labs, Inc.*

Solt, S. (2016). On measurement and quantification: The case of most and more than half. *Language*, *92*(1), 65-100. https://doi.org/DOI 10.1353/lan.2016.0016

Sparks, L., Villagran, M. M., Parker-Raley, J., & Cunningham, C. B. (2007). A patient-centered approach to breaking bad news: Communication guidelines for health care providers. *Journal of Applied Communication Research*, *35*(2), 177-196. https://doi.org/10.1080/00909880701262997

Spencer-Oatey, H., & Kádár, D. Z. (2021). *Intercultural politeness: Managing relations across cultures*. Cambridge University Press.

Stavropoulou, C. (2011). Non-adherence to medication and doctor–patient relationship: Evidence from a European survey. *Patient education and counseling*, *83*(1), 7-13. https://doi.org/10.1016/j.pec.2010.04.039

Stewart, A. J., Wood, J. S., Le-Luan, E., Yao, B., & Haigh, M. (2018). 'It's hard to write a good article': The online comprehension of excuses as indirect replies. *Quarterly Journal of Experimental Psychology*, *71*(6), 1265-1269. https://doi.org/10.1080/17470218.2017.1327546

Street Jr, R. L., Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician–patient communication to health outcomes. *Patient education and counseling*, *74*(3), 295-301.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Bott, L., Adams, R., Venetis, C. A., Whelan, L., Hughes, B., & Chambers, C. D. (2016). Exaggerations and caveats in press releases and health-related science news. *PloS one*, *11*(12), e0168217. https://doi.org/10.1371/journal.pone.0168217

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., Ogden, J., Whelan, L., Hughes, B., & Dalton, B. (2014). The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Bmj*, *349*. https://doi.org/ARTN g701510.1136/bmj.g7015

Szabolcsi, A. (2010). *Quantification*. Cambridge University Press.

Szymanik, J., & Zajenkowski, M. (2010). Comprehension of simple quantifiers: Empirical evaluation of a computational model. *Cognitive science*, *34*(3), 521-532. https://doi.org/10.1111/j.1551-6709.2009.01078.x

Tesser, A., Rosen, S., & Tesser, M. (1971). On the reluctance to communicate undesirable messages (the MUM effect): A field study. *Psychological Reports*, *29*(2), 651-654.

Touvier, M., Méjean, C., Kesse-Guyot, E., Pollet, C., Malon, A., Castetbon, K., & Hercberg, S. (2010). Comparison between web-based and paper versions of a self-administered anthropometric questionnaire. *European journal of epidemiology*, *25*(5), 287-296. https://doi.org/10.1007/s10654-010-9433-9

Tracy, K. (1990). The many faces of facework.

Van Noorden, R. (2015). Interdisciplinary research by the numbers. *Nature*, *525*(7569), 306-307. https://doi.org/10.1038/525306a

VandeKieft, G. (2001). Breaking bad news. *American family physician*, *64*(12), 1975. https://www.ncbi.nlm.nih.gov/pubmed/11775763

Vaske, J. J. (2008). *Survey research and analysis: Applications in parks, recreation, and human dimensions*. Venture Publishing, Incorporated.

Vergis, N., & Terkourafi, M. (2015). The role of the speaker's emotional state in im/politeness assessments. *Journal of language and social psychology*, *34*(3), 316-342. https://doi.org/10.1177/0261927x14556817

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of business research*, *88*, 428-436. https://doi.org/10.1016/j.jbusres.2017.12.043

Warnes, G. R., Bolker, B., Lumley, T., Warnes, M. G. R., & Imports, M. A. S. S. (2018). Package 'gmodels'. *Vienna: R Foundation for Statistical Computing*.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 781. https://doi.org/Doi 10.1037/0096-1523.16.4.781

Weinstein, Y., & Sumeracki, M. A. (2017). Are twitter and blogs important tools for the modern psychological scientist? *Perspectives on Psychological Science*, *12*(6), 1171-1175. https://doi.org/10.1177/1745691617712266

Westerståhl, D. (1984). Determiners and context sets. *Generalized quantifiers in natural language*(4), 46-71.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of experimental psychology: General*, *143*(5), 2020. https://doi.org/10.1037/xge0000014

Wiseman, H., Chappell, P., Toerien, M., Shaw, R., Duncan, R., & Reuber, M. (2016). Do patients want choice? An observational study of neurology consultations. *Patient education and counseling*, *99*(7), 1170-1178.

Woelfle, M., Olliaro, P., & Todd, M. H. (2011). Open science is a research accelerator. *Nature chemistry*, *3*(10), 745-748. https://doi.org/10.1038/nchem.1149

Wood, L. A., & Kroger, R. O. (1991). Politeness and forms of address. *Journal of language and social psychology*, *10*(3), 145-168.

Woodfield, H. (2008). Interlanguage requests: A contrastive study. *Developing contrastive pragmatics: Interlanguage and cross-cultural perspectives*, 231-264.

Woodzicka, J. A., & LaFrance, M. (2001). Real versus imagined gender harassment. *Journal of Social Issues*, *57*(1), 15-30. https://doi.org/Doi 10.1111/0022-4537.00199

Zajenkowski, M., & Szymanik, J. (2013). Most intelligent people are accurate and some fast people are intelligent.: Intelligence, working memory, and semantic processing of

quantifiers from a computational perspective. *Intelligence*, *41*(5), 456-466. https://doi.org/10.1016/j.intell.2013.06.020

Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. *Journal of psycholinguistic research*, *43*(6), 839-853. https://doi.org/10.1007/s10936-013-9281-3

Zikmund-Fisher, B. J. (2011). Time to retire the 1-in-X risk format. In (Vol. 31, pp. 703-704): Sage Publications Sage CA: Los Angeles, CA.

Zikmund-Fisher, B. J. (2014). Continued use of 1-in-X risk communications is a systemic problem. In (Vol. 34, pp. 412-413): Sage Publications Sage CA: Los Angeles, CA.

Zolnierek, K. B. H., & DiMatteo, M. R. (2009). Physician communication and patient adherence to treatment: a meta-analysis. *Medical care*, *47*(8), 826. https://doi.org/10.1097/MLR.0b013e31819a5acc