The face of research: Do first impressions based on the facial appearance of scientists affect the selection and evaluation of science communication?

Ana I. Gheorghiu

A thesis submitted for the degree of Doctor of Philosophy

Department of Psychology

University of Essex

July 2017

Table of Contents

ACKNOWLEDGEMENTS

First of all, I would like to thank the University of Essex, the Psychology Department and the ESRC for offering me the opportunity to pursue a PhD in a topic that I love. I am also thankful to Dr. Marie Juanchich and Prof. Peter Hegarty, who have been kind enough to examine my thesis, and provide invaluable feedback and advice.

Out of all the amazing people I have met throughout the course of my studies, I am most indebted to my supervisors, Dr. Will Skylark and Prof. Mitch Callan. Being supervised by two incredibly gifted and dedicated academics has taught me the value of hard work, resilience, following strict research practices and maintaining a positive attitude even when it was hard doing so. Dr. Skylark and Prof. Callan have done so much more than just share their knowledge, expertise and insight. They have taught me life lessons which I will carry with me in my career as a researcher. I am eternally grateful to my supervisors for guiding me, teaching me the delicate balance between working hard and enjoying life, and the importance of supporting your fellow researchers through difficult times, academic or personal. I would not be the researcher I am today, nor the person I am today without them, and words cannot express my respect and admiration for them. Thank you for being my supervisors, my mentors, and mostly importantly, thank you for being my friends when I needed it.

I would also like to thank all the amazing people in the Psychology Department for helping and supporting me throughout the years and for making my experience a truly unforgettable one. I have always felt at home here, and I owe it all to you.

Finally, I want to extend a heartfelt thank you to the most important people in my life: my partner, my friends and my family, who have all been incredibly supportive and have responded to my mood swings and breakdowns with unconditional love. I love you with all my heart, and I would not have done this without you.

ABSTRACT

First impressions based on facial appearance alone predict a large number of important social outcomes in areas of interest to the general public, such as politics, justice and economics. The current project aims to expand these findings to science communication, investigating both the impressions that the public forms of a scientist based on their facial appearance, and the impact that these impressions may have on the public's selection and evaluation of the research conducted by the scientist in question. First, we investigated what social judgement traits predict looking like a "good scientist" (someone who does high-quality research) and an "interesting scientist" (someone whose research people show interest in). Three studies showed that looking competent and moral were positively related to both looking like a good scientist and to interest ratings, whereas looking physically attractive positively predicted being perceived as a scientist with higher interest ratings, but was negatively related to looking like a good scientist. Subsequently, we investigated whether these perceptions translated into real-life consequences. Three studies examined the impact of first impressions on the public's choice of scientific communications, and found that people were more likely to choose real science news stories to read or watch when they were paired with scientists high on interest judgements. Another three studies looked at whether the appearance of the researcher influenced people's evaluations of real science news stories. We found that people judged the research to be of higher quality when it was associated with "good" scientists. Our findings illustrate novel insights into the social psychology of science communication, and flag a potential source of bias in the dissemination of scientific findings to the general public, stemming solely from the facial appearance of the scientist.

AUTHOR'S NOTE

Some of the research presented here has been published (Gheorghiu, Callan & Skylark, 2017). This thesis replicates some of the structure and content of our publication.

CHAPTER 1: PROJECT BACKGROUND

Science communication and miscommunication

The influence of scientific research on public policy, government issues, and domains which impact the general public is increasing (e.g., neuroscientific findings have been involved in public policy development, Seymour & Vlaev, 2012). Evidence based practices are also being developed in areas like medicine, nursing, public health and social work (Satterfield et al., 2009). For example, behavioural science theories have been used to inform public health and health-promotion interventions (Glanz & Bishop, 2010).

Considering the involvement of scientific research in areas that affect the general public, the importance of a clear and straightforward communication between scientists and the public cannot be emphasised enough. Knowledge that has been transmitted fluidly from scientists to the citizens, usually via the press (Logan, 2001) is referred to as "science communication"; this has been seen as a way of informing the public about relevant information, regarding events and developments in the scientific community (Treise & Weigold, 2002). Ideally, the wider public should be informed, and have a clear understanding of the scientific research presented, prior to making an educated decision regarding policies informed by scientific progress (Hartz & Chappell, 1997). Because the news and media are often used as an intermediary between the scientist and the public, a large number of people have started using information relayed by the press as their main source of scientific information, without comparing it with the original scientific articles (Nelkin, 1995).

Controversially, recent evidence has indicated that people are highly influenced by online commenters (particularly when said commenters are perceived to be credible), when discussing health-related public service announcements (Kareklas, Muehling & Weber, 2015). This increased reliance on the media as the main source of scientific information, along with the potential biases that people are exposed to, can pose the risk of

miscommunication in terms of scientific findings. Such was the case for the combined measles, mumps and rubella (MMR) vaccine, an example of scientific miscommunication with severe repercussions in today's society: the original report claiming that the MMR vaccine causes autism (Wakefield et al., 1998), was given too much credibility by the media (Moore, 2006). The claims have since been retracted, discredited and proved to be fraudulent (Flaherty, 2011), but parents are still reluctant to have their children vaccinated, reducing herd immunity and increasing the chance of an epidemic. Thus, scientific miscommunications can pose real and severe problems, especially when involved in matters of public importance (e.g., health).

Although measures have been taken to inform and familiarise the public with scientific articles (e.g., Das. 2013), the underlying reasons behind ineffective science communication need to be addressed. Treise and Weigold (2002) argued that ineffective communication could be triggered by either the people performing the communication (e.g., scientists assuming their audience has substantial background knowledge of the subject, or journalists editing the communication without sufficient knowledge of the topic at hand) or by an inefficient communication process (e.g., using a catchy and impactful title which does not accurately reflect the content of the communication, due to editorial pressures or journalistic targets and deadlines). The lack of clarity regarding the exact factors contributing to scientific miscommunication, and its potential consequences, have elicited a focus on improving science communication over the past years.

As a response, scientists have made efforts to convey their research to the public in an accessible, yet clear and accurate manner. Initially, communication improvements were aimed at the structure and the content of the scientific message; for example, Hartley (2003) argued that structured journal abstracts (e.g., with separate subtitles and sections) were to be preferred to traditional abstracts, due to their increased clarity and ease of read.

Furthermore, using a writing style appropriate for the reader's level of knowledge on the topic, and including less jargon in the text have been suggested to lead to clearer communication (Kirkman & Turk, 2002). The context of the scientific communication has also been found to influence how the information is perceived: Corbett and Durfee (2004) presented participants with one of two versions of a news story about global warming, and then asked them to report how certain they felt about global warming occurring. Participants who read the news story set in a scientific context reported feeling more certain that global warming was occurring, whereas participants who read the news story set in a journalistic/controversy context reported feeling less certain (Corbett & Durfee, 2004).

Besides language and content, argument and text structure can also increase clarity in communication; for example, extracting the relevant information, and presenting it in the order of importance for the reader has been a successful method of improving communication (Savić, 2003). Shonkoff and Bales (2011) investigated both modifications at the level of the content, and at the level of the argument structure in a large group project designed to clearly explain the science of child development to policymakers and the general public. During the project, neuroscientists, developmental psychologists and communication experts worked as a team, and successfully informed the public of complex scientific issues, using simplified concepts, metaphors and argument structures based on simple storylines (Shonkoff & Bales, 2011). Scientific communication could also be improved on a higher-level, by improving institutional frameworks, designed to support good communication. For example, Illes et al. (2010) suggested a scheme designed to improve neuroscience communication, comprising of: (a) additional support for the development of communication experts, (b) rewards for reaching out and communicating with a wider audience, and (c) more support for continuing research into improvement of scientific communication. Thus, methods of improving scientific communication have been centred

on text and argument structure clarity, as well as around ways to enable and encourage communication (see Illes et al., 2010).

Socio-cognitive perspectives: Predictions based on persuasion models

The improvements discussed above focus on the linguistic side of science communication, on the message itself; however, one could also approach the social psychology side of science communication, aiming to understand perceptions of scientific messages and how people form beliefs about aspects of the world. Considering that judgements are not always based on effortful, conscious information processing, and can simply rely on an automatic, intuitive process (Kahneman, 2011), it is crucial to examine the social and psychological factors potentially influencing the public's opinion of scientific communications.

Findings from domains with important social implications, such as politics, law and academia, suggest that the public is capable of forming accurate impressions of people from photos or short videos. Such inferences have been found to have predictive power in terms of outcomes (e.g., election outcomes, criminal sentences and course ratings; Ambady, Bernieri & Richeson, 2000). Visual media, such as widely available YouTube videos, TED talks and podcasts, is highly accessible and popular on the Internet as an easy way of gaining scientific knowledge, unlike traditional journal articles, which may be less accessible or easy to comprehend. Since visual media is increasingly used in distributing novel scientific findings to the public, it would prove fruitful to investigate how the appearance of a scientist may influence the public's perception of scientific research.

Additional support for the hypothesis that the facial appearance of a scientist could potentially play a role in shaping people's beliefs about the research itself comes from the Elaboration Likelihood Model (ELM, Petty & Cacioppo, 1986), which suggests that there different routes to persuasion: central and peripheral. Similarly, the Heuristic-Systematic

model of information processing (HSM; Chaiken & Trope, 1999) argues that individuals use both systematic processing and heuristic processing (e.g., relying on superficial cues, such as physical appearance) when judging argument quality. In line with this hypothesis, Lenz and Lawson (2011) found evidence that people relied more on the appearance of a political candidate when they had less knowledge of politics, and increased exposure to television, reinforcing the possibility that the appearance of a scientist may influence the public's opinion of scientific research, especially for those with low engagement in science. Connor and Siegrist (2010) investigated whether having more knowledge substantially influenced the public's perceptions of risks or benefits of gene technology. Their results suggested that social trust (Siegrist & Cvetkovich, 2000) in institutions regulating gene technology, rather than knowledge about the topic, predicted how likely members of the public were to accept gene technology (GMO; Connor & Siegrist, 2010). Thus, investigating the predictive and diagnostic social psychological information that may affect how the public judges scientific research could provide valuable information regarding science communication.

This project aims to approach science communication from a novel perspective, by addressing the social psychology side of the problem. I wish to approach this underexplored facet of science communication by investigating the effects of first impressions built on the scientist's facial appearance. More precisely, my goal is to clarify what perceptions of physical appearance define a credible scientist, whether these perceptions influence people's beliefs about the research presented, and how to improve science communication based on these findings.

Impression formation: the effects of first impressions on real life outcomes

People form global first impressions of a person by bringing together individual pieces of information about them (Uleman & Kressel, 2013). Since first impressions have predictive

power in domains such as education and politics, it is reasonable to assume that first impressions may also influence the public's perception of scientific messages. To illustrate this point within the area of education, a classic study provided evidence that end-of-semester student evaluations of teachers can be predicted from the initial impressions naïve participants formed from short videos of the teachers (Ambady & Rosenthal, 1993). Participants saw three 10s silent videos of each teacher, and had to rate the teachers on various personality dimensions; the ratings were shown to reliably predict end-of-semester student evaluations of the teachers (Ambady & Rosenthal, 1993; replicated in Ambady & Gray, 2002, Study 1). The results replicated even when shorter videos - 5s or 2s - were used to create the initial impressions (Ambady & Rosenthal, 1993). Such effects are not limited to education: participants correctly judged the relationship between two people (strangers, friends, or romantically involved) after viewing a 15s silent video of the pair interacting (Ambady & Gray, 2002). Both judgements of teacher effectiveness and relationship type were impaired by mood (sadness, in particular, reduced judgement accuracy; Ambady & Gray, 2002). To conclude, research into impression formation has shown that people can create accurate impressions of strangers from brief visual displays, and that such impressions have predictive power in terms of related outcomes.

Impression formation outside science communication: effects in politics, finance and law

First impressions have been shown to have predictive power in other areas of life as well, and an impressive amount of research has focussed on the effects of impressions based on facial appearance on electoral success: Todorov, Mandisodza, Goren and Hall (2005) found that US senate election outcomes could be predicted by people's ratings of how competent the US congressional candidates appeared, as judged from their photos. The study reported that 71.6% of the elections were won by the more competent-looking candidate; the competence judgements were based purely on first impressions, after excluding participants

who recognised candidates (Todorov et al., 2005). Moreover, Rule and Ambady (2010) asked participants to rate the faces of electoral candidates, and found that candidates higher in competence and lower in warmth were more likely to win an election. A similar line of research showed that looking more competent increased a candidate's chances of winning the election, while candidates who looked more sociable had less chances of winning an election (Castelli, Carraro, Ghitti & Pastore, 2009). Interestingly, participants in Castelli et al.'s (2009) study believed that candidates perceived to be high in both competence and sociability had higher chances of winning an election; however, this effect did not translate into actual election outcomes - only competence was a predictor of electoral success. Mattes et al. (2010) found that more threatening-looking candidates were not more likely to win, and that competence and attractiveness yielded opposite effects (higher competence, and lower attractiveness led to electoral success). Poutvaara, Jordahl and Berggren (2009) found that babyfacedness (i.e., looking low on facial maturity) was negatively correlated with perceived competence, but not related to actual electoral success; beauty was the strongest predictor of electoral success for women, while for men, it was perceived competence (Poutvaara et al., 2009). Joo, Steen and Zhu (2015) incorporated feature analysis, trait prediction and election outcome prediction to create trained models that could classify outcomes of major political events using photos of the candidates only, with over 60% accuracy. Joo et al. (2015) found that older participants were perceived as looking more competent, and that different traits predicted winning Governor races (favouring confident, attractive, energetic and masculine candidates), as opposed to Senatorial races, that favoured old, rich and competent candidates.

Further research has shown that financial decisions can also be affected by first impressions: unfakeable facial features signalling trustworthiness lead to higher investments in financial trust games (Rezlescu, Duchaine, Olivola & Chater, 2012). Trustworthy faces attracted more money when participants made judgements based on facial appearance alone, and this

effect remained even when participants had additional reputational information about the target faces, suggesting that the effect of facial appearance can survive rich environments where more information is available (Rezlescu et al., 2012). On a larger scale, companies with powerful looking CEOs (here "powerful" was defined as being rated high on looking competent, facially mature and dominant) were found to be more financially successful, suggesting that facial appearance has predictive power over economic outcomes (Rule & Ambady, 2008). Looking low on facial maturity leads to impressions of kindness, warmth, weakness, honesty and naivety; therefore, in situations of PR/financial crisis, babyfaced CEOs are more likely to be believed when denying wrongdoings of the company (Gorn, Jiang & Johar, 2008). Gorn et al. (2008) have provided evidence for the reverse effect as well: in situations where innocence is a liability, more mature-faced CEOs were the preferred choice for a new CEO, indicating the importance of situational context on the effect of face-based impressions.

Lastly, even judicial outcomes have been predicted from impressions formed based on facial appearance: babyfaced defendants were more likely to win cases that involved intentional actions, but more likely to lose cases of negligence in small claims courts (Zebrowitz & McDonald, 1991). Further research has shown that convicted murderers with more stereotypically African-American facial features had higher chances of receiving the death penalty, compared to their peers (Eberhardt, Davies, Purdie-Vaughns & Johnson, 2006; please note the statistical analysis conducted by Francis, 2015, discussing the success rates of Eberhardt et al.'s studies). Similar results were found by Blair, Judd and Chapleau (2004), who found evidence that targets with more Afrocentric features received harsher criminal sentences; the authors hypothesized that the effect was due to stereotypes about Black Americans. There was a small effect of race, but the effect of Afrocentric features was found for both White and Black participants, when examined within each group (Blair et al., 2004)

The studies described above suggest there is a certain element of universality regarding impression formation, since its effects are present in numerous, distinct areas. Thus, we reasonably expect to find similar effects of character evaluation, when scientific communications include visual depictions of the scientist.

Thin slices of life

As previously exemplified, the impressions people form within a few hundred milliseconds from another person's physical appearance can be both accurate and reliable (Olivola & Todorov, 2010a). Given the high reliability of these judgements, the current project will employ a "thin slices of life" methodology, which refers to the use of photos, or short, silent videos of a person (either performing an action, and engaging in social interaction), as representations of expressive behaviour (Ambady et al., 2000). Visual displays that do not include audio information are often used when investigating impression formation effects (e.g., silent videos and photos, as opposed to videos containing speech and audio information), considering that nonverbal information is more accessible and easier to process for the person viewing the visual communication (Ambady et al., 2000). Dynamic displays (videos) convey higher levels of information than static displays (photos); the more nonverbal information one has about the target they are rating, the more accurate their judgements will be (Weisbuch & Ambady, 2011). Such judgements based on nonverbal cues have been found to be accurate, and to influence subsequent judgements about the target's actions; in turn, these judgements correlate with real-life outcomes (Naylor, 2007). Judgement accuracy increases if the personality traits that are being judged are considered important for the domain the target is being judged on (Naylor, 2007).

Research so far has not reached a consensus regarding the ideal length of a "thin-slice" video enabling people to form accurate first impressions, with videos used in research varying widely in length (e.g., 2 seconds; Ambady & Rosenthal, 1993; 12 minutes; Borkenau,

Mauer, Riemann, Spinath & Angleitner, 2004). However, there are arguments suggesting that people can form accurate first impressions from a 10s "thin-slices" video (Weisbuch & Ambady, 2011), that 60s videos provide an appropriate compromise between video duration and judgement accuracy (Carney, Colvin & Hall, 2007) and that thin-slices videos should not be longer than 5 minutes (Ambady, LaPlante & Johnson, 2001). Finally, a quantitative review of 30 studies found that facial expressions of emotion were judged equally accurately both from short (1s) exposures, and longer exposures, suggesting the exact duration of the thin-slices may not be crucial (Hall, Andrzejewski, Murphy, Schmid Mast & Feinstein, 2008). Recent research has found promising results regarding the reliability and validity of thin slices: amongst other findings, Murphy et al. (2015) suggested that 30s to 1 minute slices reliably represent the behaviour they depict, increasing the confidence in using thin-slices methodology in research. A comprehensive review by Todorov, Olivola, Dotsch and Mende-Siedlecki (2015) argued that 34ms is enough to form an impression, and that there is no difference in accuracy of impressions above 200ms. Overall, the more information transmitted by the thing slice (i.e., longer exposure, dynamic and audio information), the more reliable the judgements will be; however, in the interest of understanding the way impressions are formed in the real world, we will attempt to recreate the first impressions formed from little information, and in a short time span.

The type of images or videos used as thin slices in the impression formation literature has varied across time, from computer-generated images of faces, that can be manipulated on the various dimensions of interest (e.g., Oosterhof & Todorov, 2008), to real-life photos of people. Advanced manipulation techniques allowed researchers to determine which information contained in photos was relevant to social judgements: Dotsch and Todorov (2012) used reverse correlation to extract psychologically meaningful images that map onto social perception, and found that certain regions of the face (i.e., the mouth, eyes, eyebrows and hair regions) contained the most diagnostic information. For example, faces judged to

look trustworthy had larger eyes and a smiling mouth, while faces judged to look more dominant had strong eyebrows, and a slightly downturned mouth (Dotsch & Todorov, 2012). These findings were supported by Todorov et al. (2015), who suggested angry faces are seen as more dominant, while smiling faces are seen as more trustworthy. More recently, there has been evidence supporting the use of ecologically valid stimuli, as opposed to the classic, computer generated faces: Sutherland et al. (2013) used ambient images to generate a 3D face model, and argued that the natural variation in face stimuli in the real world is no reason to strictly control stimuli. Their 3D face models found very similar social judgement dimensions (trustworthiness, dominance and youthful-attractiveness), suggesting the use of ecologically valid stimuli could be as accurate as computer-generated faces (Sutherland et al., 2013). Vernon, Sutherland, Young and Hartley (2014) provided additional evidence for this claim, illustrating that, despite huge variation in ecologically valid photos, a good amount of the variance in first impressions was accounted by changes in defined features. The researchers were able to build a model (from ecologically valid stimuli) that predicted first impressions of ambient images not seen by the model before, thus generalising its performance to untrained faces (Vernon et al., 2014). Thus, our plan for this project is to maintain as much ecological validity as possible, by using real-life photos of scientists.

In sum, research suggests that social decisions are shaped by first impressions, which are, in turn, shaped by facial appearance (Olivola, Funk & Todorov, 2014). We aim to use thin-slices methodology to investigate what first impressions people form based on visual information about a scientist, and whether this can affect people's opinion of the scientific message.

Models of social judgement

2-factor model of social judgement (competence and warmth)

There are two fundamental dimensions of social judgement: competence and warmth (Judd, James-Hawkins, Yzerbyt & Kashima, 2005). Different research groups consider slightly different, but very similar dimensions to be fundamental in social judgement: dominance and trustworthiness (Oosterhof & Todorov, 2008), competence and warmth (Fiske, Cuddy, Glick & Xu, 2002; Fiske, Cuddy & Glick, 2007), competence and morality (Wojciszke, 1994), agency and communion (Abele & Wojciszke, 2007); all these findings suggest that a 2-factor model is best for investigating social judgements. There are numerous examples of the 2-factor model in the impressions formation literature; for example, Ambady, Krabbenhoft and Hogan (2006) asked participants to rate sales managers on interpersonal (e.g., empathic, warm, collaborative) and task-oriented (e.g., achieving, persevering and task-oriented) traits, after listening to 20s audio clips of the managers speaking. Interpersonal (warmth) ratings were more strongly correlated with each other than with task-oriented (competence) measures and vice-versa, highlighting the warmth-competence space of social judgements (Ambady et al., 2006). In line with this model of social judgement, traits such as efficacy, intelligence and ability correlated with the dimension of competence, whereas traits like kindness, likeability and trustworthiness correlated with the dimension of warmth (Fiske et al., 2007). Dominance, competence and facial maturity have also been linked with the dimension of power, while trustworthiness and likeability were linked with warmth, in a similar dichotomous division (warmth-competence vs. warmth-power) of social judgement dimensions (Rule et al., 2010).

Walker and Vetter (2016) investigated whether the BIG2 (communion and agency) and the BIG5 (neuroticism, extraversion, openness to experience, agreeableness and conscientiousness) would map onto trustworthiness and dominance. Although the BIG5 did

not have a strong overlap with the two dimensions (agreeableness and openness to experience overlapped with trustworthiness, but there was no other strong correlation), both dimensions of the BIG2 mapped well onto dominance (strong correlation with agency) and trustworthiness (strong correlation with communion; Walker & Vetter, 2016). These results reinforce the similarities between the many combinations of dimensions that form the 2-factor model of social judgement (competence/dominance/agency vs. warmth/trustworthiness/communion), and support our decision to focus on competence and warmth for this project.

Originally, the 2 factors were considered to be orthogonal: Oosterhof and Todorov (2008) used Principle Components Analysis on trait judgements of neutral faces, and found two orthogonal dimensions, which they labelled as valence/trustworthiness and dominance. Valence was found to be more sensitive to avoidance/approach signals, while dominance was more sensitive to strength/weakness signals, suggesting there is some adaptive significance for these facial cues (Oosterhof & Todorov, 2008). Although these two dimensions appear to be fundamental for social judgements, Oosterhof and Todorov (2008) argue that the situational context can render other dimensions of face evaluation more important. More recent research has found evidence that the BIG2 of social perception may not be as orthogonal as initially proposed: Imhoff and Koch (2017) argued that the relationship between agency and communion is actually curvilinear. The researchers looked at previously published, as well as novel data, and propose the impressions of communion are highest for individuals average on agency, and that individuals low or high on agency will be perceived as lower on communion. The curvilinear relationship would explain why studies have found either a negative, orthogonal, or positive relationship between agency and communion, depending on the level of agency of their sample. Ultimately, Imhoff and Koch (2017) argue that people cannot be perceived as both highly agentic and highly communal.

Regardless of the exact relationship between the dimensions, the two-factor model of social judgement (warmth and competence) appears to be the classic view in the impression formation literature.

3-factor model of social judgement (competence, warmth and sociability)

Conversely, a new strand of research has found evidence suggesting that a three-factor model of social judgement may be more appropriate, where the warmth dimension is seen as encompassing two different concepts: sociability and morality (e.g., Heflick, Goldenberg, Cooper & Puvia, 2011). Historically, warmth has been considered to capture several aspects, such as morality and trustworthiness (Cuddy, Fiske & Glick, 2008), as well as sociability and friendliness (Kervyn, Bergsieker & Fiske, 2012). Because sociability and morality are seen as two different social concepts, the notion of warmth is likely to create conceptual ambiguity, by encompassing elements of both (Goodwin, 2015); therefore, Goodwin (2015) suggested investigating sociability and morality as two different dimensions, as opposed to being placed under the umbrella term of "warmth".

In support of the 3-factor model, Wokciske and Klusek (1996) collected ratings of people's overall approval or disapproval of their country's president's ability to handle his job. Up to 95% of people's negative impressions were predicted by the perceived competence, sociability and morality of the president (Wojciske & Klusek, 1996). Leach, Ellemers and Barreto (2007) investigated the importance of morality, competence and sociability in perceptions of groups, and found that a three-factor model (competence, morality, sociability) provided a better fit than a two-factor model (competence and warmth). They suggested that morality (measured by honesty, sincerity and trustworthiness), competence (measured by intelligence, competence and skill) and sociability (measured by likeability, warmth and friendliness) are distinct in-group characteristics (Leach et al., 2007). Similarly, Brambilla, Rusconi, Sacchi and Cherubini (2011) looked at the distinct and dominant role of

the morality component of warmth in an impression formation setting, as opposed to group processes (as illustrated by Leach et al., 2007). The authors found that morality (measured by sincerity, honesty, righteousness and respectfulness) and sociability (measured by kindness, friendliness, warmth, likeability and helpfulness) were processed differently, and a confirmatory factor analysis provided support for the three-factor model (Brambilla et al., 2011).

Evidence suggests that morality is processed differently to sociability (Brambilla et al., 2011). Goodwin, Piazza and Rozin (2014) take this idea further, proposing that one's overall impression (positive or negative) of another individual is more strongly predicted by the target's perceived morality, than the target's perceived warmth or competence. Goodwin et al. (2014) found that overall impressions of targets were best predicted by information regarding the moral character of the target, regardless of whether the targets were real or hypothetical. This suggested that moral character could be one of the most important sources of information used in impression formation. In a review of the emerging literature on the unique contribution of morality and sociability to social judgement, Brambilla and Leach (2014) provided additional evidence that morality has a primary role over sociability; for example, information regarding the morality of a stranger affected participants' first impressions of them more than information regarding non-moral characteristics (Pagliaro, Brambilla, Sacchi, D'Angelo & Ellemers, 2013). This effect was also found in intergroup impressions: when reading about an unfamiliar outgroup, participants' first impression was mostly affected by whether the outgroup was described as high or low on morality (Brambilla, Sacchi, Rusconi, Cherubini & Yzerbyt, 2012). Interestingly, while morality characteristics seem to be more important when forming global impressions of other people, this is not the case for self-perception and self-attitudes—evaluations of the self are influenced more by our own competence than our own morality (Wojciszke, 2005).

The primacy of morality over other judgements may be due to evolutionary importance of detecting trustworthiness. Research indicated that trustworthiness judgements are performed faster than judgements about the target's sociability or competence (Brambilla & Leach, 2014), and neuroimaging evidence suggests that detecting trustworthiness may be an automatic process linked to the amygdala (a brain structure involved in detecting threats; Brambilla & Leach, 2014; Willis & Todorov, 2006). Goodwin (2015) proposed that morality, sociability and competence have unique contributions to forming impressions about others due to the each of the factors pointing to different aspects of other people. According to the author, morality indicates a person's intentions, competence indicates how capable the person is to carry out their intentions, while sociability indicates whether a person will be successful in recruiting allies to support their intentions (Goodwin, 2015). Landy, Piazza and Goodwin (2016) expanded on this hypothesis, and argued that, since morality indicates the nature of one's goals and competence/sociability indicate the likelihood that a person will accomplish their goals, being high on competence and sociability will be seen as positive, depending on one's perceived morality. In other words, being highly moral is unconditionally seen as positive, whereas high sociability and competence will only be regarded as positive in moral others: friendly and smart people will be disliked if their goals are immoral, since their friendliness and competence suggest they are likely to achieve their (immoral) goals (Landy et al., 2016). Results suggesting competence and sociability are positive contingent on morality support the primacy of morality judgements, as well as the unique contribution morality is likely to make to first impressions in science communication.

Therefore, we plan to use competence, morality and sociability as different dimensions describing people's first impressions of scientists, to provide additional evidence for a three-factor model of social judgement in the context of science communication.

Traits and dimensions

Competence, sociability and morality

We will focus on the 3-factor model of social judgement by incorporating the 3 main traits of competence, sociability and morality. These factors illustrate the basic dimensions on which people evaluate groups and individuals, and we argue that they are crucial to science communication. How competent a person looks has been shown to predict positive outcomes in many areas of life (e.g., Todorov et al., 2015). Scientists are sometimes depicted as somewhat incompetent, or absent-minded (Haynes, 2003). Despite this, intelligence and skill have been found to be central to both competence (Fiske et al., 2007), and to scientist stereotypes (Mead & Metraux, 1957), suggesting one should expect high competence to have a positive effect on people's perception of a scientist. Competence alone may not be enough: trust is another important element for both effective communication and to the scientific process (e.g., Godfrey-Smith, 2003; Shapin, 1996). Fiske and Dupree (2014) found that perceptions of scientists included both warmth and competence (in particular high competence but low warmth), arguing that communicators need both expertise and trust. In support for this line of research, evidence suggests that trustworthy-looking scientists may enjoy greater research success (Dilger, Lutkenhoner & Muller, 2015). Even though morality has been shown to have little effect in other areas where trust is important (e.g., politics, Mattes et al., 2010), it is worth considering its effect on science communication. Even though science communication is a social endeavour, scientists themselves are often seen as rather solitary and socially-awkward (Schinske, Cardenas & Kaliangara, 2015). While appearing sociable seems to be a desirable quality in those communicating science in a school environment (Mendez & Mendez, 2016), being too sociable may have the opposite effect on a scientist, diminishing their appearance as a "good scientist", and in turn reducing the public's regard for their work (Martinez-Conde,

2016). Thus, this work suggests science communication should be examined by taking into account all these facets that might play a role in people's perception of scientists.

Attractiveness

Besides the three core socio-cognitive traits, we will consider the possibility that facial attractiveness may influence first impressions of a scientist. Attractiveness has been found to create a halo effect: attractive people tend to be rated highly on other positive traits, such as intelligence and social competence (Miller, 1970; Eagly, Ashmore, Makhijani & Longo, 1991). Sofer, Dotsch, Wigboldus and Todorov (2015) investigated how face typicality (i.e., attractiveness) may affect trustworthiness judgements, and found that familiar faces were liked more, and were considered to be "safer" (i.e., more trustworthy). Perceived trustworthiness decreased as faces moved away from the typical face; although trustworthiness judgements correlated with attractiveness judgements, as faces became more typical, trustworthiness judgements followed a positive trend, while attractiveness judgements followed a negative one (Sofer et al., 2015). In terms of science communication, it is not clear how attractiveness may influence inferences made from the scientist's facial appearance. Although attractiveness is valued in communicators (e.g., Mendez & Mendez, 2016), it does not predict research success (Dilger et al., 2015). Talamas, Mavor and Perret (2016) investigated the effects of attractiveness on perceived and actual academic performance in the classroom, thus focussing on the students rather than the educator. Their results show a stronger effect of the attractiveness halo in perceptions of female intelligence. Furthermore, there was no consensus on the relation between attractiveness and actual academic performance; however, there was a positive relationship between attractiveness and perceived academic performance (Talamas et al., 2016). By looking at the effects of attractiveness in other areas, one might expect attractiveness to have a detrimental effect on people taking a scientist's work seriously (e.g., Mattes et al., 2010).

Gender differences

Throughout our research, we aimed to consistently address and consider the issue of gender differences in scientific research and social dimensions. Gender differences have been a long-standing problem in science, and recent work from the impression formation literature shows that the issue is still salient.

There is evidence suggesting that, when looking at the implicit associations made between gender and science, people regard scientists as male by default. These implicit stereotypes predicted gender differences in scientific and mathematical school achievement (Nosek et al., 2009). Knobloch-Westerwick, Glynn and Huge (2013) discussed the "Matilda effect" – the under-recognition of female scientists, and argued that the preference for male authors may be a consequence of conservative gender norms. Agentic roles are more often associated with men, while communal roles are associated with women. Given the individualistic nature of a scientist's role and the communal qualities of women, this apparent mismatch may lead to higher perceived scientific quality and more collaboration interest for male authors in male topics (Knobloch-Westerwick et al., 2013). These stereotypes have been found to translate into real-life consequences, with evidence of gender bias in faculty hiring decisions: in a large-scale study, male applicants were perceived as more competent and hireable than female applicants; the women were seen as less hireable due to being perceived as less competent (Moss-Racusin, Dovidio, Brescoll, Graham & Handelsman, 2012). Furthermore, a feminine appearance continues to be perceived as a signal that women are not well suited for science (Banchefsky, Westfall, Park & Judd, 2016).

Contrary to this line of research, advances in terms of equality illustrate a positive shift in faculty preference for women in the STEM field, in the context of hypothetical hiring experiments across a number of scientific domains (Williams & Ceci, 2015). Moreover, double-blind peer review has been found to favour an increase in female authored papers

(Budden et al., 2008). Considering the lack of consistency in the evidence, and the rather opposing results, we will examine the potential effects of a scientist's gender on impression formation, to address this issue. We will also discuss gender differences in the social traits (i.e., warmth and competence) and areas of science (i.e., biology and physics) to be used in our studies.

Gender differences in scientific areas

Throughout our experiments, we used scientists and research articles from both biology and physics, to investigate the subtle difference between a life-oriented and a more abstract side of science. When considering the stereotypes regarding differences in cognition between genders, it would not be surprising if there were gender differences in terms of preference for different areas of science. When asked to rate a list of role names on their gender stereotypicality, chemists and physicists were perceived as predominantly male (>60% male; Gabriel, Gygax, Sarrasin, Garnham & Oakhill, 2008), whereas women appeared to be overrepresented in caring occupations such as nursing, or rehabilitation therapy (McLean & Kalin, 1994). This trend seems to emerge during school years: both boys and girls perceive physical science and technology-related courses as appropriate for boys, and life sciences as appropriate for girls; biology is favoured by girls and physics is favoured by boys - these patterns persist through college and graduate school (Farenga & Joyce, 1999).

Gender differences in social dimensions

Differences in the gender stereotypes that people hold have been found both in cognition traits (i.e., rationality and mathematical reasoning seen as masculine traits, while intuition and creativity seen as feminine traits) and personality traits (i.e., masculine personalities seen as competitive, aggressive, dominant, etc., with feminine personalities considered as sympathetic, sensitive, warm; Diekman & Eagly, 2000, Diekman & Goodfriend, 2006). The above-mentioned differences in how masculine and feminine personalities are perceived

map well on the warmth/communion and competence/agency dimensions of social judgement: women are judged to be more communal (warm, caring) and less agentic (assertive, competent) than men (Fuegen, Biernat, Haines & Deaux, 2004). In turn, these differences are also reflected in the roles and jobs associated with each gender: women's traditional gender roles emphasize communal over agentic traits (Fiske, Cuddy & Glick, 2007), and women can even internalise these discrepancies (e.g., despite a lack of difference in actual performance, young girls rated their own math competency lower than did boys; Herbert & Stipek, 2005). Furthermore, women are typically less respected (illustrating competence) or liked (illustrating warmth), but not both: prejudice exists against both women who choose a traditional path (perceived as incompetent, but warm) and women who are career-orientated (perceived as competent, but cold; Cuddy, Fiske & Glick, 2004, Fiske, Xu, Cuddy & Glick, 1999). Although warmth is stereotypically feminine, women are perceived as less warm, competent and moral when people focus on their appearance (Heflick et al., 2011). Given that these gender effects are linked to the dimensions of social judgement that we will be considering in our research, it is important to look into any potential gender differences that might emerge.

"Good" scientists vs. "Interesting" scientists

Throughout this project, we aimed to tap into two key components of science communication, which we expect to be influenced by a scientist's facial appearance, in the context of the strength of face-based first impressions, and of the susceptibility of the science communication to superficial factors. Namely, we investigated both the process of selection (i.e., which research the public chooses to read) and of evaluation (i.e., the opinions the public forms about the research they read). To this end, we created two measures, tapping into different sides of being a scientist: being a good scientist (i.e., someone who conducts good research, following the scientific method), and being an "interesting" scientist (i.e., someone whose research people would be interested in finding

more about). There is some evidence in the literature towards these two facets of being a scientist: Martinez-Conde (2016) found that popular, visible scientists were perceived by their peers to be worse academics than those who do not engage with the public, potentially due to a worry that such scientists care more about their social media presence than about science and research. Although these perceptions did not translate into real-life consequences (scientists who engaged with the public were actually more academically active; Martinez-Conde, 2016), it appears that these two sides of science are fairly distinct, and will be investigated as such in the current project.

Summary of traits

Based on the impression formation literature, the main traits/dimensions we examined were competence (measured by competence and intelligence - based on research indicating that a target's actual intelligence can be predicted from their perceived intelligence; Murphy, Hall & Colvin, 2003), sociability (measured by likeability and kindness), morality (measured by trustworthiness and honesty) and perceived physical attractiveness. In addition, perceived age, gender and ethnicity of the scientists were measured to explore their potential contributions to the participants' ratings of their other traits (e.g., Chiao, Bowman & Gill, 2008). In sum, the first part of this project will focus on the science communicator's competence, sociability and morality, as well as their age, physical attractiveness, gender and ethnicity, aiming to reveal what people perceive as a "good scientist" and a scientist likely to garner interest in their work.

Present research

We investigated the social psychology of science communication, namely what first impressions people form based on a scientist's appearance, and whether these impressions are likely to influence the communication of scientific findings to the general public. This is of particular importance in the context of an increased use of visual media to transmit

scientific messages to the general public. The project aimed to determine what social traits determine the image of a "good scientist" and of a scientist likely to garner interest in their work, and whether these two concepts differ. Additionally, we investigated whether looking like a good scientist or one with high interest ratings will imprint a positive image onto the research the respective scientists are transmitting, by testing the public's choice and opinions of the scientific messages. This was achieved using both rating studies (aimed at discovering the traits/dimensions important in defining a good scientist and a scientist likely to be perceived as interesting) and validation studies (aimed at investigating whether these impressions have a real-life effect on how people perceive and select scientific messages). The key research questions were: 1) Which traits/dimensions are most important in defining a "good" scientist, and a scientist likely to produce interesting research, and do these dimensions differ? 2) Do impressions created by facial appearance predict science communication outcomes? 3) Are the effects large enough to be of practical significance as a potential source of bias? Based on previous research, facial competence, morality, sociability and attractiveness are plausible influences on the public's selection and evaluation of scientific communications; however, the direct and magnitude of these effects are open questions that the current project aims to address. We will also investigate the possibility that any such effects will be stronger for participants with little engagement with science, as they may rely more on superficial cues (e.g., appearance; Petty & Cacioppo, 1986).

CHAPTER 2: THE FACE OF RESEARCH. WHAT SOCIAL TRAITS DEFINE SCIENTISTS? DO "GOOD" SCIENTISTS DIFFER FROM SCIENTISTS LIKELY TO PRODUCE INTERESTING RESEARCH?

Study 1: What social traits define scientists, from photos of US scientists? Do "good" scientists differ from scientists likely to produce interesting research?

Study 1 was designed to investigate what characteristics (i.e., dimensions) define someone who looks like a good scientist, and someone who looks likely to produce interesting research (from here on referred to as "interest judgements" or "interesting research"). Participants rated a large set of photos of scientists on various dimensions. We used these ratings to test the three-factor model of social judgement (Brambilla et al., 2011).

Method

1. Participants

All participants were recruited from the University of Essex pool of volunteers. The sample size for this study was based on previous impression-formation studies, which have found reliable results with 20-25 participants per stimulus (e.g., Willis & Todorov, 2006). One participant was excluded for having zero variance in their trustworthiness ratings, suggesting little to no engagement with the task.

For the task of rating scientist faces on the predictor variables, 53 participants (9 men and 44 women) took part, with ages ranging from 18 to 50 (M = 20.0, SD = 4.6). Approximately 51% reported being British nationals and having English as their first language.

For ratings of the scientist faces on the criterion variables, 54 participants (16 men and 38 women) were recruited, and received £3 for their time. Ages ranged from 18 to 40 (M = 21.4, SD = 4.9), and all participants were British nationals (94% also had English as their first language).

2. Stimuli and Materials

The stimuli for both stages were photos of scientists, collected from US University websites, and were either photos of geneticists (selected from departments of 'Genetics' or 'Human Genetics') or physicists (selected from departments of 'Physics'). Only photos from the main faculty were selected (Lecturers, Professors, Assistant, Associate and Emeritus Professors were selected, as opposed to administrative staff and technicians), in the interest of collecting stimuli representative for the image of a scientist. Initially, a set of photos of geneticists was collected as follows: Universities were randomly selected, one at a time, from the 200 top-ranked US Universities (National University Rankings, 2014). If the selected University had a Genetics/Human Genetics department, 10 photos of faculty members were randomly selected from each University's departmental website. The initial set of geneticists' photos comprised 254 photos (some Universities had a low number of faculty members or photos of members, resulting in fewer than 10 photos). The sampling procedure was based on an estimated number required to produce a final stimuli sample of 100-120 photos (after editing). For the purpose of the current study, this set has been supplemented by exhaustively searching all of the remaining Universities to identify those with Genetics/Human Genetics departments, resulting in an additional 10 photos. The current image set therefore comprises randomly-sampled faces from all of the top-200 US Universities that have Genetics/Human Genetics departments. The photos were edited so that the face of the person appeared on a grey background (RGB coordinates 124 123 123; colour code #7b7c7c). The same grey was used as the background colour for the experimental display, so that only the actual face of the person stood out. The images were cropped, so that the final image commenced at the top of the person's head, and finished immediately below the chin; the sides were also cropped, to ensure that the face was as centred as possible. Finally, any images that were too small (below 130 pixels in height) were removed, and the remaining images were resized to have a height of 130 pixels; any poor-quality images were excluded, resulting in a final stimuli set of 108 photos of

geneticists (103 from the initial set, and an additional 5 following the exhaustive search). The sampling of physicists followed the same process (only recruiting photos from Physics departments) and aimed to produce the same number of useable images. (Note that because Physics departments are more common than Genetics departments, this sample did not exhaustively represent the top-200 Universities with Physics departments). The initial set for physicists comprised 271 photos, which produced a final sample of 108 photos of physicists.

The experiment was run using PsychoPy v1.81.00 (Peirce, 2007). Participants completed the experiment in individual testing booths. The study was run on 21.5 inch screen iMacs running at a screen resolution of 1920 x 1080.

3. Design and Procedure

For both stages, participants were presented with photos of scientists and asked to rate each scientist's face on various dimensions.

In stage 1 (measuring ratings for the predictor variables), each participant saw half the photos, and was asked to rate each photo on the following dimensions/traits: competence, intelligence, likability, kindness, trustworthiness and honesty. Ratings of the scientists' physical attractiveness and perceived age were also collected from the same participants. The ratings were done in eight blocks, one type of judgement per block (e.g., competence). All 108 photos were presented in each block, and their order was randomised separately in each block. The order of the blocks was also randomised for each participant. On-screen instructions (e.g., "How competent is this person?"; "How old is this person?") prompted participants for their response for each trial, appearing above the respective photo. All judgements (except for ratings of age) were made on a Likert-type scale ranging from 1 (Not at All) to 9 (Extremely), using the keys on top of the keyboard. The button-press terminated

the presentation of the image, and the next stimulus followed after a 500ms delay. For age, participants typed in the perceived age using the same keys, and their answer appeared underneath the image; participants could also modify their initial answer using the Backspace key, and confirm their answer using the Enter/Return key. In total, participants made 864 judgements each, equivalent to 8 blocks of 108 stimuli.

Since each participant only saw half the photos, two versions of the task were created, each presenting half of the stimuli (54 photos of geneticists and 54 photos of physicists, randomly allocated to the two versions). Participants' allocation to either version 1 or 2 of the study was counterbalanced.

For stage 2 (measuring ratings for the criterion variables), each participant saw all the photos. Half of the participants were asked to indicate how likely it was that each person was a good scientist (i.e., that they conduct accurate scientific research which yields valid and important conclusions; task 1), whereas the other half were asked to indicate how interested they would be to find out more about the person's research (task 2). All 216 photos were presented in a single block, and their order was randomised for each participant. On-screen instructions (e.g., "How likely is it that this person is a good scientist?" or "How interested would you be in finding out more about this person's research?") prompted participants for their response for each trial, appearing above the respective photo. The same response scales and methodology from stage 1 were replicated. In total, participants made 216 judgements each; their allocation to either task 1 or 2 of the study was counterbalanced.

Demographic information was collected at the beginning of both stages, alongside measures of participants' engagement with science. The latter were obtained by using a novel questionnaire, touching on aspects of people's knowledge of science (e.g., "I am knowledgeable about science", "I fully understand the scientific method"), and their interest

in science (e.g., “I find scientific ideas fascinating”, “I have little interest in science”; see Appendix A).

Ratings of gender (male or female) and ethnicity (Caucasian or non-Caucasian) were obtained by collecting ratings from two independent judges, and resolving disagreements with ratings from a third judge.

Results

1. Data preparation

The predictor data was prepared as follows: any age ratings where the participant had pressed ‘Return’ without entering a number were coded as zero and excluded. Any age judgements below 16 or over 100 were also flagged and removed from the data set. Mean judgement ratings for each task-face combination were computed. The criterion data was rearranged by calculating the mean ratings for each task-face combination. Please note that from here onwards, dichotomous variables were coded as follows for all subsequent analyses: male = 0, female = 1 (gender); White = 0, non-White = 1 (ethnicity); biology = 0, physics = 1 (discipline); 0 = text, 1 = video (format); low = 0, high = 1 (face-type).

2. Internal Reliability

To assess the internal reliability of the scales used in both stage 1 and stage 2, two Cronbach’s Alpha values were calculated for each faceset-dimension combination[1] in stage 1, and for each task in stage 2: an Alpha value robust against non-normality and missing data (package *coefficientalpha* for R; Zhang & Yuan, 2015), and a non-robust value (equivalent to the calculations used by SPSS). Although all scales had very good to excellent internal reliability (see Table 1), the “interesting research” measure had the lowest

[1] Faceset here refers to the two different sets of stimuli used in Version 1 and Version 2 of stage 1. This precaution was necessary since each participant only saw half of the total number of faces (either faceset 1 in version 1, or faceset 2 in version 2).

reliability, suggesting that people have different perceptions of what someone who conducts interesting research looks like, or even of what research they would find interesting.

Measure	Faceset	Cronbach's Alpha	
		Robust	Non-robust
Competence	1	0.86	0.86
	2	0.83	0.84
Likeability	1	0.90	0.90
	2	0.91	0.91
Trustworthiness	1	0.86	0.86
	2	0.88	0.89
Intelligence	1	0.86	0.86
	2	0.89	0.89
Kindness	1	0.90	0.90
	2	0.93	0.93
Honesty	1	0.88	0.88
	2	0.89	0.89
Attractiveness	1	0.93	0.94
	2	0.93	0.95
Age	1	0.99	0.98
	2	0.99	0.99
Good scientist	-	0.89	0.89
Interesting research	-	0.73	0.72

Table 1. Robust and non-robust values of Cronbach's Alpha for the eight predictors and two criterion variables ("Good scientist" and "Interesting research") in Study 1.

3. Science Engagement Questionnaire

Because we created the Science Engagement Questionnaire (see Appendix A), a Principal Components Analysis (PCA) was conducted on the science questions, to investigate its underlying structure. With PCA, a single factor was automatically extracted with an Eigenvalue above 1, which accounted for 58% of the total variance. The scree plot did not suggest that more than one factor would be appropriate in this case, and the questions had good loadings onto the extracted factor. The suitability of the analysis was reinforced by two measures of appropriateness of factor analysis, a Keiser-Meyer-Olkin Measure of Sampling Adequacy test (.888) and Bartlett's Test of Sphericity ($\chi^2(28) = 453.3$. $p < .001$), both returning satisfactory results. The internal reliability of the one-factor scale was calculated, resulting in very good internal reliability (Cronbach's $\alpha = .89$ for science engagement). Finally, the composite measure was computed by averaging the ratings across all eight items, and labelling the final variable Science Engagement.

4. Confirmatory Factor Analysis

Considering the theoretical background underpinning the current study, a confirmatory factor analysis (CFA) was conducted, to confirm that competence, sociability and morality are the three underlying factors (as opposed to the classic two-factor model of competence and warmth; e.g., Fiske et al., 2007). Within the analysis, each item was loaded only onto its hypothesized factor, and the latent variables were allowed to correlate with each other. We used the *lavaan* package for R (Rosseel, Oberski & Byrnes, 2011) to conduct the CFA.

Correlations among the trait ratings verified that they are consistent with the three-factor structure we were expecting (see Table 2).

	M (SD)	Intelligent	Likeable	Kind	Trustworthy	Honest
Competent	5.78 (0.78)	0.860*	0.385*	0.404*	0.543*	0.522*
Intelligent	5.79 (0.86)		0.310*	0.376*	0.515*	0.511*
Likeable	5.13 (1.01)			0.914*	0.799*	0.850*
Kind	5.41 (1.09)				0.850*	0.904*
Trustworthy	5.06 (0.94)					0.903*
Honest	5.38 (0.96)					

*Table 2. Correlations among the items forming each trait for Study 1 (mean, SD; * indicates p <.05).*

In the three-factor model, competence and intelligence loaded onto Competence, likeability and kindness loaded into Sociability, while trustworthiness and honesty loaded onto Morality. In the two-factor model, the Competence factor remained the same, while likeability, kindness, trustworthiness and honest loaded onto a single factor labelled Warmth. Maximum estimation likelihood was employed to estimate all models. After investigating global indices of fit, the three-factor model (SRMR = .018, RMSEA = .102, CFI = .991, TLI = .978, BIC = 2056.26) was a better fit than the two-factor model (SRMR = .051, RMSEA = .247, CFI = .931, TLI = .870, BIC = 2139.70), as indicated by a chi-square test for the difference in model fit: χ^2diff(2) = 94.19, p<.001. The three-factor model also fit better than a single-factor model tapping into the overall impression: SRMR = .142, RMSEA = .419, CFI = .776, TLI = .627, BIC = 2370.55; χ^2diff(3) = 330.41, p<.001. The results suggest that competence, sociability and morality are the underlying factors, providing evidence against a classic two-factor model of competence and warmth. The final model is illustrated in Figure 1, where circles represent latent variables, and rectangles represent measured variables.

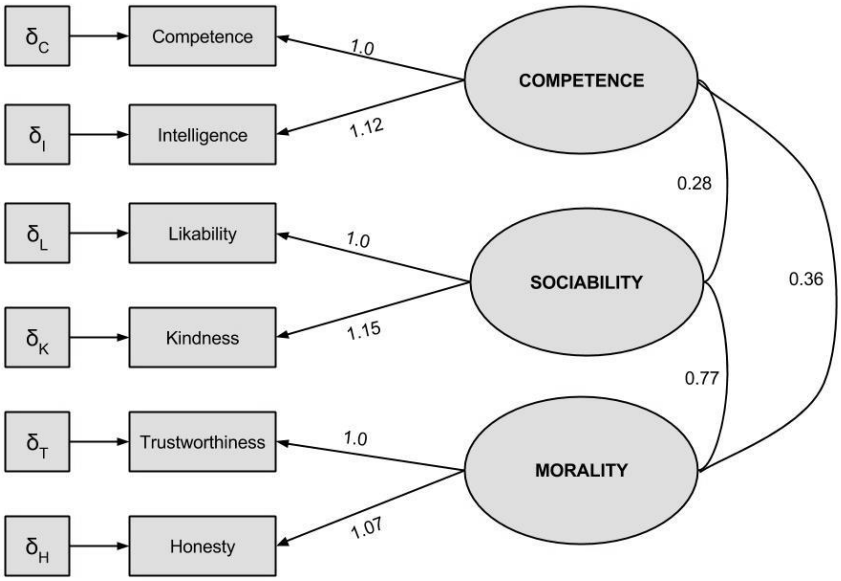


Figure 1. Final CFA model for Study 1 with significant, unstandardized coefficients.

Composite traits were calculated by averaging across the corresponding items, according to the CFA. Correlations between the composite traits and the criterion variables were computed, illustrating the trends one might expect to see in the data, such as a strong correlation between morality and sociability, and between competence and looking like a good scientist (Table 3).

	M (SD)	Face Ethnicity	Face Age	Attract.	Comp.	Sociability	Morality	Good Scientist	Interesting Research
Face Gender	0.20 (0.40)	0.127	-0.169*	0.481*	0.141*	0.396*	0.478*	-0.091	0.260*
Face Ethnicity	0.16 (0.37)		-0.196*	-0.021	0.144*	0.005	0.080	0.173*	-0.018
Face Age	46.23 (9.81)			-0.590*	0.110	-0.191*	-0.092	0.451*	-0.269*
Attractiveness	3.61 (1.11)				0.231*	0.563*	0.489*	-0.248*	0.696*
Competence	5.78 (0.79)					0.390*	0.555*	0.778*	0.505*
Sociability	5.27 (1.03)						0.893*	0.098	0.632*
Morality	5.22 (0.93)							0.304*	0.624*
Good Scientist	5.62 (0.91)								0.182*
Interesting Research	4.84 (0.59)								

*Table 3. Correlations between the composite traits and the criterion variables in Study 1 (mean, SD; * indicates p <.05).*

5. Mixed Effects Models

Finally, the data from both the predictor ratings and the criterion ratings were collated, and analysed to assess the impact of the predictors (competence, morality, sociability, collapsed from the collected ratings as a results of the CFA and attractiveness) and controls (age, gender and ethnicity) on to the criterion variables (good scientist and interesting research).

Due to the design of the experiments, the data can be analysed either at the level of the face, by averaging across ratings for each face (without taking into account that different people saw different faces) or by fitting a mixed effects model, and taking into consideration that participants performed one task of the other. Both options were investigated and results were compared.

a. Face-level data

Firstly, face-level data was analysed, checking for clustering by University – due to collecting more than one photo per University, we were able to test whether a nested model provided a better fit. For both 'good scientist' and 'interesting research' separately, two models were fitted using the *lme4* package (Bates, Maechler, Bolker & Walker, 2015): the null, constant-only model and a model containing only a random intercept for University. The two models were compared using a custom-made test employing the log-likelihood ratio, and the results suggested that the models did not differ for either looking like a good scientist ($\chi^2(1) < .001$, $p = .999$) or for looking likely to produce interesting research ($\chi^2(1) = 3.63$, $p = .057$). Since the analyses do not indicate that face-type is systematically clustered within Universities, it is reasonable to proceed with a linear approach.

Using the *lme4* (Bates et al., 2015) and *lmerTest* packages (Kuznetsova, Brockhoff & Christensen, 2015), we predicted looking like a good/interesting scientist from the fixed

effects of the scientist's age, gender, ethnicity and discipline, and ratings of attractiveness, competence, sociability and morality (no random effects). As shown in Table 4, age, gender ethnicity, attractiveness, competence, sociability and morality predicted looking like a good scientist, while gender, attractiveness, competence and morality (borderline) predicted looking like a scientist likely to produce interesting research.

Predictor	Good Scientist		Interesting Research	
	B	*p*	B	*p*
Age	0.177	< .001	0.047	.166
Gender	-0.068	.048	-0.115	< .001
Ethnicity	0.079	.007	-0.009	.717
Attractiveness	-0.252	< .001	0.374	< .001
Discipline	0.040	.164	0.012	.616
Competence	0.698	< .001	0.136	< .001
Sociability	-0.152	.024	0.059	.330
Morality	0.204	.007	0.124	.065

Table 4. Coefficients and p-values for all the fixed effects of the predictors (age, gender, ethnicity, attractiveness, discipline, competence, sociability and morality), when using face-level data to predict 'looking like a good scientist' and 'looking likely to produce interesting research' in Study 1.

To confirm that there is no effect of University, two more models were built for both good scientist and interesting research: a random intercept model (fixed effects: discipline, gender, ethnicity, age, attractiveness, competence, sociability and morality; random effects: random intercept for University) and a random intercept + slopes model (fixed effects: discipline, gender, ethnicity, age, attractiveness, competence, sociability and morality; random effects: random intercept for University and random uncorrelated slopes for the by-discipline, by-gender, by-ethnicity, by-age, by-attractiveness, by-competence, by-sociability

and by-morality effect of University). After observing the additional variance explained by adding random effects of University, it appears that the random effects do not explain a large amount of additional variance. The three models were compared using the same log-likelihood ratio custom test, and the results indicated that the models did not significantly differ for either looking like a good scientist or looking likely to produce interesting research (see Table 5). Thus, there was no particular benefit to adding random effects for University in our models.

Test	Good Scientist		Interesting Research	
	χ^2	*p*	χ^2	*p*
A vs. B	7.67	.467	0.950	.999
B vs. C	< .001	.999	0.440	.507
A vs. C	7.67	.568	1.39	.998

Table 5. Chi-square values and p-values for the comparison between the three models fitted to data in Study 1 for both 'looking like a good scientist' and 'looking likely to produce interesting research', where A = Model with random intercepts and random slopes, B = Model with random intercepts only and C = Model with no random effects.

After investigating the data at the level of the face, it seems to be the case that scientists who looked older, male, not Caucasian, not perceived as attractive, competent, unsociable but moral were more likely to be judged as looking like a good scientist. Similarly, scientists who looked male, attractive and competent were judged as more likely to produce interesting research. Two separate analyses indicated that there was no effect of the University from which the photos were collected.

b. Mixed effects data

To investigate the data using a mixed effects design, all the data points were used, without performing any averaging across faces. In addition to the previously analysed dimensions,

information about the participants was also attached to the dataset (participant's age, gender and engagement with science score).

Using *lme4* (Bates et al., 2015), we built two models for both good scientist and interesting research: a random intercepts model (fixed effects: scientist's age, gender, ethnicity, attractiveness, discipline, competence, sociability and morality, and the participant's age, gender and science engagement; random effects: random intercepts for each participant and each scientist) and a random slopes model (the same fixed effects and random intercepts, with additional random uncorrelated slopes for the by-age, by-gender, by-ethnicity, by-attractiveness, by-discipline, by-competent, by-sociability and by-morality effect of participant, and for the by-participant age, by-participant gender and by-participant science engagement effect of scientist).

The random intercepts and random slopes models were compared using a likelihood-based model test, and the random slopes model provided a better fit for both looking like a good scientist ($\chi^2(11) = 458.21$, $p < .001$) and for looking likely to produce interesting research ($\chi^2(11) = 504.32$, $p < .001$). Thus, adding random slopes to our model accounts for a significant amount of variation; hence, the random slopes model will be reported.

The mixed effects model replicated the results obtained in the face-level analysis: looking like a good scientist was predicted by the scientist's age, attractiveness, competence, sociability and morality, while looking like a scientist likely to produce interesting research was predicted by the scientist's perceived attractiveness and competence (see Table 6).

Predictor	Good Scientist		Interesting Research	
	B	*p*	B	*p*
Age	0.177	.007	0.047	.446
Gender	-0.068	.143	-0.115	.084
Ethnicity	0.079	.179	-0.009	.792
Attractiveness	-0.252	< .001	0.374	< .001
Discipline	0.039	.160	0.013	.613
Competence	0.698	< .001	0.136	.026
Sociability	-0.152	.023	0.059	.496
Morality	0.204	.012	0.124	.068
Participant age	-0.247	.138	0.030	.872
Participant gender	-0.099	.548	-0.169	.409
Participant science engagement	0.026	.877	0.382	.073

Table 6. Coefficients and p-values for the fixed effects of the predictors (age, gender, ethnicity, attractiveness, discipline, competence, sociability, morality, participant age, gender and science engagement) for the random intercept and slopes model, when using mixed-effects data to predict 'looking like a good scientist' and 'looking likely to produce interesting research' in Study 1.

After investigating the mixed-effects structure of the data, the results suggest that scientists who looked older, not perceived as attractive, competent, unsociable but moral were more likely to be judged as looking like a good scientist. Conversely, scientists who looked more attractive and competent were judged as more likely to produce interesting research.

Discussion

The results of Study 1 confirmed that a three-factor model of social judgement (competence, morality and sociability; Brambilla et al., 2011) provided a better fit for our

current data, as illustrated by the confirmatory factor analysis. More importantly, Study 1 has helped define the image of a "good scientist" as someone older, not attractive, competent, moral but unsociable, when the random variation between participants was taken into account. Similarly, when random effects are taken into account, the image of a scientist likely to produce interesting research was defined by someone looking attractive and competent. Study 1 helped highlight the importance of traits such as competence, sociability and morality in defining the stereotypical image of scientist, but also illustrated the opposing effect of attractiveness in terms of "good" scientists versus "interesting" scientist. The next step involved validating the results by investigating whether these differences in appearance have a real-life effect on people's perception of scientific research.

Study 2: What social traits define scientists, from photos of members of the public? Do "good" scientists differ from scientists likely to produce interesting research?

The aim of Study 2 was to replicate the effects found in Study 1 using different stimuli and a largely non-University sample of participants. In Study 1 the stimuli were photos of scientists, randomly selected from University websites. Due to the naturalistic character of the stimuli, the sample was unbalanced in terms of gender (there were few female scientists). By using photos from a face database in Study 2, we could verify whether the same effects emerged in a more balanced sample of photos of average people rather than scientists. Finally, a "scientist" measure was added to investigate whether people distinguish looking like a scientist, looking like a good scientist and looking likely to produce interesting research.

Method

1. *Participants*

The sample size was based on obtaining 25-30 ratings for each dimension; participants were recruited online, using Amazon's MTurk. Two participants who rated the predictor variables and one participant who rated the criterion variables were eliminated for having zero variance in their data, suggesting a lack of engagement with the task.

For the predictor ratings, 60 participants took part during the initial batch of testing. However, after an initial analysis, it was decided to collect data from more participants in order to achieve more accurate ratings, and reduce sampling variation: 20 more useable participants were recruited, ten for each face-set. The final sample of 80 participants was comprised of 41 men and 39 women, with age ranging from 20 to 60 years old (M = 34.4, SD = 10.9). All participants were American, and approximately 93% had English as their first language.

For the criterion ratings, 91 participants (58 men) were tested, with age ranging between 19 and 64 (M = 32.9, SD = 9.3). Ninety-six percent of participants had English as their first language, and all participants were American.

2. *Stimuli and Materials*

The stimuli for both stages were photos collected from the Park Aging Mind Face Database (Minear & Park, 2004). Images from the "Neutral Faces" database were used, and a final sample of 200 photos was created by randomly selecting 50 faces from each age group (18-29; 30-49; 50-69 and 70-94 years old). Out of the 50 faces selected from each age group, half were photos of women, and half were photos of men, resulting in a balanced sample in terms of both the age and gender of the people in the photos. Any photos where the hair or

face were obstructed (e.g., people wearing a headscarf) were removed, and another photo was randomly selected to replace it. Additionally, the ethnic mix of the selected sample was designed to represent the ethnic mix of the UK (to illustrate a sample of UK scientists, which would be less likely to be recognised by US participants), according to the 2011 Census data (Office for National Statistics, 2013): approximately 85-90% of the people depicted in the photos were White, with the remaining 10-15% being Non-White (percentages depending on the number of available photos in each age group). The original photos were not edited or cropped in any way, preserving the original standardised dimensions and white/off-white backgrounds.

3. Design and Procedure

Participants viewed multiple photos of scientists and had to rate them on various social judgement dimensions. For the first stage (predictor ratings), each participant saw half of the photos (100 images), and had to rate the people depicted in the photos on the same social dimensions as in Study 1. Information about the person's age was included in the face database, so perceived age ratings were not collected. The presentation and randomisation of photos was identical to the methodology in Study 1. In total, each participant made 700 judgements throughout the study.

An attention check was also employed which asked participants to choose which dimension they were not asked to rate in the study. Furthermore, participants were asked to provide details about any interruptions or breaks that might have occurred during the study, and were asked whether they had previously completed or started the study. Demographic information about the participant was collected at the end of the survey, as well as information regarding their involvement with science (see Appendix A).

In the interest of time, and of maintaining the participants' attention and interest, only half of the photos were presented to the participants. In order to ensure that participants saw and rated the same photos on all dimensions, two versions of the task were created by dividing the initial 200 photos in half, through random selection. Participants were then randomly allocated to one of the two versions.

For the second stage (criterion ratings), participants saw all 200 photos and had to rate them all on one of the following three dimensions: "scientist" ("How much does this person look like a SCIENTIST?"), "good scientist" ("How much does this person look like a GOOD SCIENTIST?") and "interesting research" ("How INTERESTED would you be in finding out more about this person's research?"); for the latter two dimensions, participants were asked to assume that the people they were about to see were scientists. Because each participant saw a single block containing all the photos, only the order in which the photos were presented was randomised. The remaining procedure was the same as in stage 1, excluding the attention check regarding the dimension not presented in the study.

Results

1. *Data Preparation*

Mean judgement ratings for each face-block combination were computed.

2. *Internal Reliability*

A non-robust Cronbach's Alpha, and a Cronbach's Alpha robust against non-normality and missing data assessed the internal reliability of the dimensions. The two values were calculated for each combination of version and dimension, and all scales appear to have excellent internal reliability (Table 7).

Measure	Faceset	Cronbach's Alpha	
		Robust	Non-robust
Competence	1	0.95	0.95
	2	0.94	0.94
Likeability	1	0.94	0.94
	2	0.95	0.95
Trustworthiness	1	0.94	0.94
	2	0.93	0.93
Intelligence	1	0.95	0.95
	2	0.94	0.94
Kindness	1	0.95	0.95
	2	0.94	0.94
Honesty	1	0.93	0.93
	2	0.93	0.93
Attractiveness	1	0.97	0.97
	2	0.97	0.97
Scientist	-	0.93	0.93
Good scientist	-	0.92	0.92
Interesting research	-	0.61	0.61

Table 7. Robust and non-robust values of Cronbach's Alpha for the seven predictors and three criterion variables ("Scientist", "Good scientist" and "Interesting research") in Study 2.

3. Confirmatory Factor Analysis

In line with the theoretical background of the study, and the findings of Study 1 (which confirmed the three-factor model as being a more suitable fit than the two-factor model), we conducted a confirmatory factor analysis (CFA) to test whether competence, sociability and morality are the three underlying factors. The *lavaan* package for R (Rosseel et al.,

2011) was used, each item was only loaded onto its hypothesized factor, and latent variables were allowed to correlate.

Correlations among the trait ratings were rather high overall, but showed a pattern similar to the three-factor structure we were expecting (see Table 8).

	M (SD)	Intelligent	Likeable	Kind	Trustworthy	Honest
Competent	5.37 (0.95)	0.946*	0.639*	0.525*	0.690*	0.622*
Intelligent	5.21 (0.92)		0.631*	0.537*	0.727*	0.668*
Likeable	4.98 (1.00)			0.948*	0.892*	0.868*
Kind	5.03 (1.03)				0.880*	0.883*
Trustworthy	5.11 (0.94)					0.965*
Honest	5.32 (0.93)					

*Table 8. Correlations among the items forming each trait for Study 2 (mean, SD; * indicates p <.05).*

Two models were computed, using a Maximum Estimation Likelihood (ML) approach: a three-factor model comprised of competence (competence and intelligence), sociability (likeability and kindness) and morality (trustworthiness and honesty), and a two-factor model comprised of competence (competence and intelligence) and warmth (likeability, kindness, trustworthiness and honesty). The two-factor model had a less-than-ideal fit for the data (SRMR = .043, RMSEA = .385, CFI = .881, TLI = .777, BIC = 1605.772); however, the three-factor model produced a negative variance for intelligence and was impossible to fit. Additional ratings for each face on all the predictors were collected in order to reduce sampling variation, and produce more accurate estimates. Unfortunately, this did not solve the problem, and the three-factor model was not fitted when using ML.

Since ML is particularly prone to model fitting problems, we changed the estimation procedure to Generalised Least Squares (GLS), and attempted to fit the three-factor model again. With a GLS estimator, the three-factor model fit the data (SRMR = .106, RMSEA = .195, CFI = .816, TLI = .541), and did so better than the two-factor model (SRMR = .098, RMSEA = .258, CFI = .571, TLI = .197). Because the model fitting was problematic and did not suggest a straightforward solution, subsequent analyses were conducted for both the two-factor and three-factor models.

Composite traits were calculated by averaging across the corresponding items, according to the CFA. Correlations between the composite traits and the criterion variables were computed, illustrating the trends one might expect to see in the data, such as a strong correlation between morality and sociability, and between competence and looking like a good scientist (Table 9).

	M (SD)	Face Ethn.	Face Age	Attract.	Comp.	Soc.	Moral.	Warmth	Good Scient.	Int Scient.	Scientist
Face Gender	0.50 (0.50)	0.025	-0.008	0.045	-0.086	0.060	0.141*	0.102	-0.294*	-0.106	-0.252*
Face Ethn.	0.19 (0.39)		-0.139	0.113	0.079	0.166*	0.107	0.141*	-0.051	0.153*	0.036
Face Age	50.49 (22.13)			-0.681*	-0.173*	-0.022	0.193*	0.083	-0.134	0.042	-0.161*
Attract.	3.92 (1.22)				0.683*	0.527*	0.415*	0.485*	0.503*	0.508*	0.481*
Comp.	5.29 (0.92)					0.598*	0.692*	0.660*	0.823*	0.736*	0.814*
Soc.	5.00 (1.00)						0.901*	0.977*	0.408*	0.632*	0.385*
Moral.	5.21 (0.92)							0.973*	0.500*	0.682*	0.473*
Warmth	5.11 (0.94)								0.463*	0.673*	0.438*
Good Scient.	4.47 (1.07)									0.730*	0.937*
Int Scient.	4.73 (0.48)										0.710*
Scientist	3.79 (1.11)										

*Table 9. Correlations between the composite traits and the criterion variables in Study 2 (mean, SD; * indicates p <.05).*

4. Mixed Effects Modelling

The data for both predictors and criterions were collated, and the following variables were computed: competence (average of competence and intelligence), sociability (average of likeability and kindness), morality (average of trustworthiness and honesty) and warmth (average of likeability, kindness, trustworthiness and honesty).

Given the structure of the data, mixed effects analyses were performed, using all the individual data points, and demographic information about the participants (age, gender and their engagement with science score). The criterion variables were regressed onto both the three-factor model of social judgement, and the two-factor model, separately, using the *lme4* package (Bates et al., 2015). We predicted looking like a scientist/good scientist/interesting research from the scientist's age, gender, ethnicity, and ratings of attractiveness, competence, sociability and morality (for the three-factor model), or ratings of attractiveness, competence and warmth (for the two-factor model). Additionally, two types of models were built for each criterion variable: a random intercepts model (comprising fixed effects specific to the three- or two-factor models, and random intercepts for each participant and each face), and a random slopes model (the same fixed effects and random intercepts, with additional random uncorrelated slopes for each by-fixed-effect effect of participant, and for each by-participant-level variable effect of face).

Likelihood-ratio tests were used to compare the random intercept-only models with their random intercept and slopes counterparts, and the more complex models provided a better fit for looking like a scientist ($\chi^2(10) = 847.12$, $p < .001$), looking like a good scientist ($\chi^2(10) = 1092.1$, $p < .001$), as well as looking likely to produce interesting research ($\chi^2(10) = 2415.4$, $p < .001$). This was also the case for models using the two-factor solution: the random

intercept and slopes models provided a significantly better fit than the random intercept models for looking like a scientist ($\chi^2(9) = 840.84$, $p < .001$), for looking like a good scientist ($\chi^2(9) = 1090$, $p < .001$) and for looking likely to produce interesting research ($\chi^2(9) = 2406.9$, $p < .001$). Hence, for all analyses, adding random slopes has accounted for a significant amount of extra variance, so we will focus on the results of the random intercept and slopes models.

When using the three-factor model, looking like a scientist was predicted by age, gender, attractiveness and competence; looking like a good scientist was predicted by gender, ethnicity and competence, while looking likely to produce interesting research was predicted by attractiveness and competence (see Table 10 and Figure 2).

	Scientist			
Predictor	B	95% CI Low	95% CI High	*p*
Age	-0.268	-0.483	-0.053	.016
Gender	-0.182	-0.346	-0.018	.034
Ethnicity	-0.028	-0.122	0.066	.557
Attractiveness	-0.355	-0.577	-0.132	.002
Competence	1.074	0.866	1.282	< .001
Sociability	-0.141	-0.390	0.109	.270
Morality	0.134	-0.188	0.455	.416
Participant age	0.073	-0.258	0.405	.667
Participant gender	-0.297	-0.651	0.057	.110
Participant science engagement	0.038	-0.326	0.401	.841

Good Scientist				
Predictor	B	95% CI Low	95% CI High	*p*
Age	-0.163	-0.402	0.076	.185
Gender	-0.249	-0.428	-0.069	.009
Ethnicity	-0.116	-0.221	-0.011	.034
Attractiveness	-0.161	-0.359	0.036	.112
Competence	0.914	0.724	1.104	< .001
Sociability	-0.204	-0.434	0.027	.085
Morality	0.233	-0.058	0.524	.119
Participant age	0.055	-0.362	0.471	.799
Participant gender	-0.243	-0.659	0.173	.262
Participant science engagement	0.417	0.0002	0.833	.060

Interesting Research				
Predictor	B	95% CI Low	95% CI High	*p*
Age	0.198	-0.047	0.442	.122
Gender	-0.048	-0.185	0.088	.492
Ethnicity	0.051	-0.028	0.131	.213
Attractiveness	0.199	0.016	0.382	.040
Competence	0.197	0.035	0.358	.022
Sociability	0.079	-0.059	0.217	.263
Morality	0.003	-0.161	0.167	.973
Participant age	0.472	-0.016	0.96	.068
Participant gender	0.279	-0.229	0.787	.290
Participant science	0.408	-0.082	0.899	.113

engagement

Table 10. Coefficients, 95%CIs and p-values for the fixed effects of the predictors (age, gender, attractiveness, discipline, competence, sociability, morality, participant age, gender and science engagement) for the random intercepts and slopes models, when using mixed-effects data to predict 'looking like a scientist', 'looking like a good scientist' and 'looking likely to produce interesting research' in Study 2.

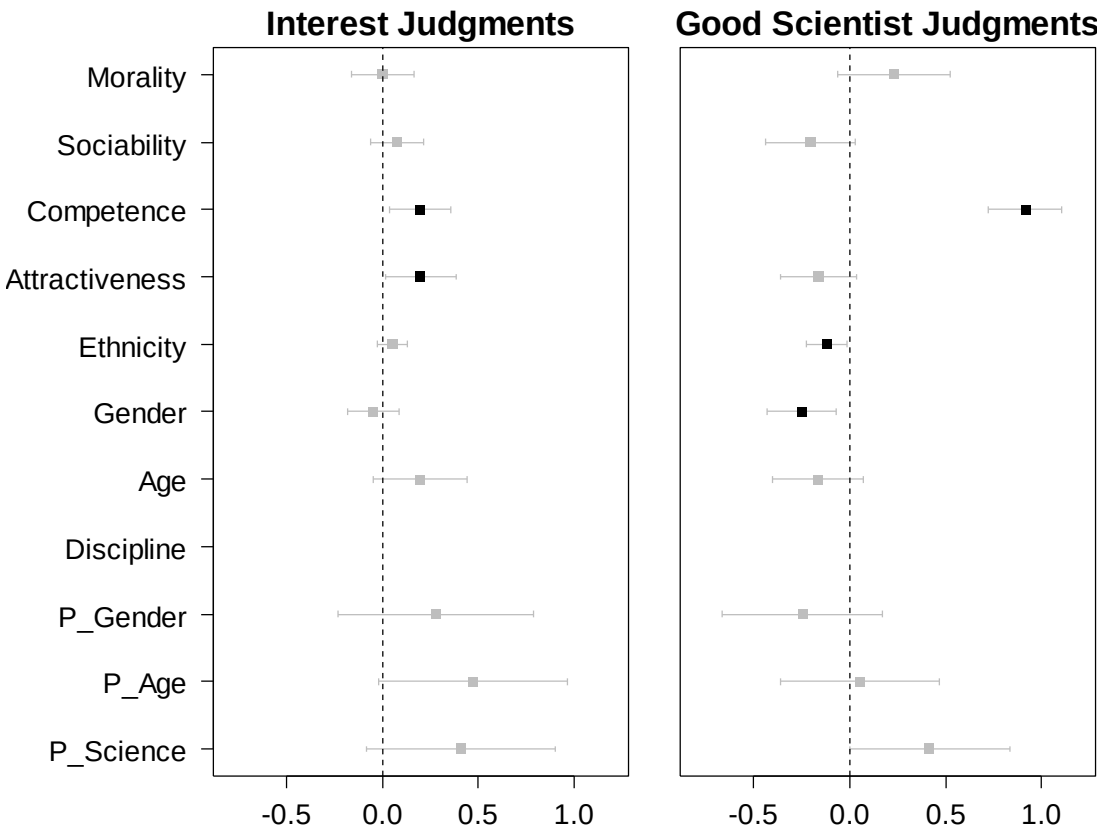


Figure 2. Regression coefficients and 95% Wald confidence intervals for the random intercept and random slopes models in Table 10. Coefficients with CIs that exclude zero are highlighted in black.

When the analyses were ran using the two-factor solution, the random slopes models showed that looking like a scientist was predicted by age, attractiveness and competence, looking like a good scientist was predicted by gender, ethnicity and competence, while looking likely to produce interesting research was predicted by attractiveness and competence (see Table 11).

Scientist				
Predictor	B	95% CI Low	95% CI High	*p*
Age	-0.237	-0.447	-0.027	.029
Gender	-0.162	-0.323	-0.002	.053
Ethnicity	-0.030	-0.124	0.063	.526
Attractiveness	-0.368	-0.589	-0.147	.001
Competence	1.124	0.946	1.302	< .001
Warmth	-0.038	-0.193	0.116	.629
Participant age	0.073	-0.258	0.405	.667
Participant gender	-0.297	-0.651	0.057	.110
Participant science engagement	0.038	-0.326	0.401	.841

Good Scientist				
Predictor	B	95% CI Low	95% CI High	*p*
Age	-0.113	-0.346	0.119	.342
Gender	-0.217	-0.393	-0.042	.020
Ethnicity	-0.119	-0.225	-0.014	.029
Attractiveness	-0.183	-0.38	0.014	.070
Competence	0.994	0.826	1.163	< .001
Warmth	-0.022	-0.158	0.115	.758
Participant age	0.055	-0.362	0.471	.799
Participant gender	-0.243	-0.659	0.173	.262
Participant science engagement	0.417	0.0002	0.833	.060

	Interesting Research			
Predictor	B	95% CI Low	95% CI High	*p*
Age	0.189	-0.053	0.432	.136
Gender	-0.054	-0.189	0.082	.444
Ethnicity	0.052	-0.028	0.132	.209
Attractiveness	0.203	0.021	0.385	.035
Competence	0.183	0.025	0.342	.030
Warmth	0.089	-0.002	0.179	.061
Participant age	0.472	-0.016	0.96	.068
Participant gender	0.279	-0.229	0.787	.290
Participant science engagement	0.408	-0.082	0.899	.113

Table 11. Coefficients,95% CIs and p-values for the fixed effects of the predictors (age, gender, attractiveness, discipline, competence, warmth, participant age, gender and science engagement) for the random intercepts and slopes model, when using mixed-effects data to predict 'looking like a scientist', 'looking like a good scientist' and 'looking likely to produce interesting research' in Study 2.

The results indicated that people judged to look more like a scientist were older, male, perceived as less attractive and more competent; people judged to look more like a good scientist were also male, Caucasian and perceived to be more competent, while people judged to look likely to produce interesting research were perceived as more attractive, and more competent.

Discussion

Study 2 complemented Study 1, indicating that looking more attractive and more competent predicted being perceived as likely to produce more interesting research, whereas looking less attractive and sociable, but more competent and moral predicted being perceived as a

better scientist. Taken together, the results of Studies 1 and 2 suggest people make a distinction between good scientists and scientists they are interested in. Core social dimensions (e.g., competence and attractiveness) play different roles in people's perceptions of good scientists and scientists likely to produce interesting research.

Study 3: What social traits define scientists, from photos of UK scientists? Do "good" scientists differ from scientists likely to produce interesting research?

Study 3 was a confirmatory study, aimed at strengthening the results of the CFAs in Studies 1 and 2, considering the difficulties we encountered with fitting a model in Study 2. The traits were slightly modified based on recent research suggesting that some of the traits we used before (e.g., kindness) loaded on both sociability and morality (Landy et al., 2016). Collecting ratings on more social dimensions for each factor (4 dimensions, as opposed to 2 dimensions in both Studies 1 and 2) should improve the model fit and achieve a more reliable estimation of the factors (Brown, 2014). Furthermore, this should allow us to select the strongest indicators for each factor, to be used in future studies (e.g., a study using thin-slice videos to form impressions of scientists). In the interest of avoiding any carryover effects (given the number of photos is not particularly large), each participant rated the photos on a single trait (Rhodes, 2006; Sutherland et al., 2013). Finally, we checked for participants' recognition of the people in the photos presented to ensure that the judgements were truly based on first impressions of the scientists (ratings were not collected for recognised faces).

Method

1. *Participants*

The required sample size was based on obtaining 25-30 judgements for each dimension, so the target sample size was approximately 780 participants for the predictor variables, and

120 for the criterion variables. The initial judgements for the criterion variables were rather low in inter-rater reliability, so the criterion sample was increased by 100 people. From here onwards all studies were conducted online, and a participant is defined as a row in the data file that had completed the task, was over 18 and was the first occurrence of that IP and MTurk ID (compared both within the study, and across other studies in this project). All participants were recruited from Amazon's Mechanical Turk, and four participants were excluded for having zero variance in their responses (suggesting a lack of involvement with the task), while another was removed due to a computer error leading to them seeing a photo more than once.

The survey was tested on an initial batch of 209 participants, to ensure the servers would not crash under the influx of data. For the predictor ratings, the final sample of 830 participants was comprised of 450 men and 380 women, with ages ranging from 18 to 72 (M = 35.3, SD = 10.8). Ninety-eight percent of the participants had English as their first language, and 97.5% were American.

For the criterion ratings, the final sample (following exclusion) consisted of 206 participants (107 men), with ages ranging from 20 to 75 (M = 34.3, SD = 10.3). Ninety-seven percent of participants were native speakers of English, while 96.6% were American.

2. Stimuli and Materials

The stimuli consisted of photos of scientists collected from Biological Sciences and Physics departments of UK university websites. A power analysis suggested that approximately 400 photos (200 from each discipline) would provide at least 85% power to detect the effects we found previously (attractiveness, competence, sociability and morality). The selection process was based on researchers submitted to the 2014 Research Excellence Framework (REF): for each unit of assessment of interest (i.e., Biological Science and Physics), we

sampled without replacement from all Universities who submitted to that unit, collecting a number of photos from each University based on the proportion of scientists from that University who submitted to the REF. For example, if the University of Cambridge contributed to 8% of the total REF submissions in the Biological Sciences unit, then we would sample 8% of the 200 Biological Sciences photos we needed from the University of Cambridge, resulting in 16 photos collected. Once the number of photos required from each University was established, the photos were collected by randomly selecting scientists from the list of names provided by the REF. If the person selected did not have a photo, the next name on the randomised list was selected. The same algorithm applied for scientists who were not employed at the University any more, those who were looking away from the camera in their photos, and those who had black and white, blurry, very low-quality photos or very small photos (thumbnails). Less commonly, people had to be removed for having a cartoon instead of a photo, for wearing sunglasses, or having their face partially covered, shaded or cut-off in the photo. Finally, if we were unable to reach the desired number of photos for a given University after exhausting the list of scientists in the REF, we would sample randomly from the whole list of scientists for that unit of assessment. However, if we were unable to get any photos from a University (and this was the case for 3 Universities), the University would be completely excluded from the total number of scientists contributing to the REF, and the proportions of photos needed from each would be recalculated, and additional photos would be samples where needed, following the same algorithm. The final sample comprised of 200 photos from Biological Sciences departments, and 200 photos from Physics departments. The photos were cropped around the top of the head and the shoulders, and standardised to 150 pixels in height, while allowing for variations in background and pose, representing the variety of photos that the public would encounter in real-life (Sutherland et al., 2013; Sutherland et al., 2015). Following standardisation, any photos that were blurry enough to make the facial features of the

scientist unclear were removed and replaced using the same sampling procedure described above, until 400 photos of acceptable quality were collected.

The study was run online, using a custom Java script. Participants completed the study in their own time, and on their own devices.

3. Design and Procedure

Participants saw photos of scientists and were asked to rate them on a social dimension. The procedure was similar to the other face rating studies described above, except participants only rated the scientists in the photos on a single social dimension, as opposed to multiple dimensions (cf. Studies 1 and 2). Participants were randomly allocated to one of the 14 social dimensions of interested: competent, intelligent, capable, effective, trustworthy, honest, moral, fair, likeable, friendly, warm and sociable, as well as perceived age and attractiveness as controls. Thus, each participant rated half of the photos (200) on a single dimension (listed above); the photos were presented one at a time, in an order randomised individually for each participant. Six different sets of 200 photos were created, and each participant was randomly assigned to one of the sets. On-screen instructions, cues and response scales were identical to those in Studies 1 and 2. The program progressed to the next photo automatically after the response was made, with a 500ms ITI and an additional 500ms of the photo being displayed without the possibility of making a response (to prevent participants from repeatedly pressing a button without viewing the photos). Participants had the option to press "r" if they recognised the person in the photo, ensuring that their judgements are *first* impressions of the scientists depicted. Demographic information about the participant was collected at the end of the survey, as well as information regarding their engagement with science (see Appendix A).

Perceived age ratings were collected separately, following the main data collection, since the program had to be re-written to allow participants to enter 2-digit numbers as their ratings. Participants pressed “Return” to submit their answer, and could use “Backspace” to edit their answers before submission.

The criterion ratings were collected after the predictors ratings were completed, and the factor structure was verified. Participants were randomly allocated to one of six sets of 200 photos, and had to rate the scientists on either how much they looked like a “good scientist” or how much they looked likely to produce interesting research (same cues as Study 1 and 2). Ratings of ethnicity were obtained per Study 1.

Results

1. Internal Reliability

Given the design of the study, it was not possible to calculate a standard Cronbach’s Alpha because the calculation excludes cases list-wise, which resulted in all cases being excluded, since participants saw one of 6 complementary sets. In order to estimate reliability, we calculated Cronbach’s Alpha for each combination of 3 non-complimentary sets, thus ensuring the function would calculate an alpha value for the common faces. All scales had good reliability averaged across all set combinations (see Table 12 for average values across all sets).

Measure	Cronbach's Alpha (average)
Age	0.99
Capable	0.74
Competent	0.78
Effective	0.72
Fair	0.75
Friendly	0.93
Good Scientist	0.89
Honest	0.81
Intelligent	0.78
Interesting Research	0.75
Likeable	0.84
Moral	0.79
Physically Attractive	0.91
Sociable	0.91
Trustworthy	0.79
Warm	0.88

Table 12. Average non-robust Cronbach's Alpha values, calculated for each dimension, across non-complementary sets of faces, in Study 3.

2. Confirmatory Factor Analysis

One of the main aims of this study was to verify the three-factor structure of competence, sociability and morality found in Study 1, considering the instability of the model in Study 2. To this end, we computed the correlations between the trait ratings to observe trends in the data (Table 13) and conducted a CFA using the lavaan package for R (Rosseel et al., 2011). Each item was only loaded onto its hypothesized factor, and latent variables were allowed to correlate.

	M (SD)	Intelligent	Capable	Effective	Likeable	Sociable	Friendly	Warm	Trust.	Honest	Moral	Fair
Competent	6.18 (0.57)	0.678*	0.715*	0.727*	0.295*	0.207*	0.155*	0.244*	0.454*	0.405*	0.459*	0.338*
Intelligent	5.69 (0.61)		0.737*	0.675*	0.123*	0.054	0.069	0.073	0.278*	0.319*	0.315*	0.135*
Capable	6.10 (0.51)			0.733*	0.210*	0.099*	0.097	0.132*	0.390*	0.404*	0.387*	0.192*
Effective	5.57 (0.58)				0.231*	0.144*	0.093	0.132*	0.361*	0.400*	0.364*	0.241*
Likeable	5.29 (0.71)					0.819*	0.806*	0.825*	0.742*	0.727*	0.729*	0.799*
Sociable	5.13 (0.96)						0.890*	0.867*	0.597*	0.612*	0.576*	0.737*
Friendly	5.30 (1.07)							0.912*	0.640*	0.692*	0.624*	0.744*
Warm	5.17 (0.87)								0.677*	0.673*	0.645*	0.753*
Trustworthy	6.07 (0.68)									0.786*	0.808*	0.709*
Honest	5.63 (0.65)										0.803*	0.702*
Moral	5.79 (0.63)											0.704*
Fair	5.47 (0.66)											

*Table 13. Correlations between the items forming each trait, for Study 3 (mean, SD; * indicates p <.05).*

Three models were computed: (1) a three-factor model comprised of competence (competent, intelligent, capable, effective), sociability (friendly, likeable, sociable, warm) and morality (fair, honest, moral, trustworthy); (2) a two-factor model where sociability and morality were collapsed into a single "warmth" factor; and (3) a one-factor model where all items loaded onto a single "overall impression" factor. All three models converged. The three-factor model (SRMR = .064, RMSEA = .128, CFI = .933, TLI = .913, BIC = 5579.77) provided a better fit of the data than the two-factor model (SRMR = .107, RMSEA = .190, CFI = .845, TLI = .808, BIC = 6004.34), or the single factor model (SRMR = .204, RMSEA = .284, CFI = .650, TLI = .572, BIC = 6969.18). Although the fit of the three-factor model is not ideal, it is still a better fit than the other models, as suggested by chi-square tests for the difference in model fit: χ^2diff(2) = 436.56, p<.001 (against the two-factor model), χ^2diff(3) = 1407.4, p<.001 (against the single-factor model). Hence, the CFA provides additional evidence that using competence, sociability and morality would be appropriate for any further analyses of our data. Thus, average variables were computed for future use, and the correlations between the composite traits and criterion variables were calculated (Table 14).

	M (SD)	Face Ethnicity	Face Age	Attract.	Comp.	Sociability	Morality	Good Scientist	Interesting Research
Face Gender	0.20 (0.40)	-0.062	-0.143*	0.395*	-0.005	0.363*	0.603*	-0.113*	0.192*
Face Ethnicity	0.07 (0.26)		-0.089	-0.161*	0.178*	-0.022	-0.003	0.233*	0.051
Face Age	43.15 (9.04)			-0.500*	0.357*	-0.123*	0.012	0.535*	0.085
Attractiveness	4.02 (0.82)				0.136*	0.330*	0.386*	-0.369*	0.442*
Competence	5.88 (0.50)					0.168*	0.424*	0.689*	0.585*
Sociability	5.22 (0.85)						0.798*	-0.069	0.422*
Morality	5.74 (0.59)							0.163*	0.534*
Good Scientist	5.79 (0.77)								0.279*
Interesting Research	4.85 (0.56)								

*Table 14. Correlations between the composite traits and the criterion variables in Study 3 (mean, SD; * indicates p <.05).*

3. Mixed Effects Models

The mixed effects analysis was conducted using participant's raw scores, and demographic information about the participants. Given the confirmatory nature of this study, we focused on a random intercepts and random slopes model, allowing us to compare the most complex models across the three face-rating studies. Furthermore, according to Barr, Levy, Scheepers and Tily (2013), random intercept only models should be avoided, in favour of maximal models, in order to avoid Type I error rate inflation. With this is mind, from here on, models with maximal but uncorrelated random effects will be used and reported in all analyses, where possible.

The model comprised of fixed effects for face-level dimensions (age, gender, ethnicity, discipline, attractiveness, competence, sociability and morality) and participant-level variables (age, gender and science engagement), as well as random intercepts for participants and faces, and random uncorrelated slopes for each by-fixed-effect effect of participant, and for each by-participant-level variable effect of face). The model was fitted twice, once for the good scientist measure, and once for the interesting research measure. The results show a similar trend to our previous face-rating studies: looking like a good scientist was predicted by looking more competent, more moral, but less attractive and less sociable; looking likely to produce interesting research was predicted by looking more attractive, more competent and more moral (see Table 15 for a side-by-side comparison between the two face-rating studies).

Good Scientist								
Predictor	Study 1				Study 3			
	B	95%CI Low	95%CI High	*p*	B	95%CI Low	95%CI High	*p*
Age	0.177	0.056	0.298	0.007	0.059	-0.019	0.137	0.140
Gender	-0.068	-0.158	0.022	0.143	0.023	-0.072	0.119	0.633
Ethnicity	0.079	-0.034	0.192	0.179	0.040	-0.014	0.094	0.146
Discipline	0.039	-0.015	0.094	0.160	0.024	-0.019	0.067	0.283
Attractiveness	-0.252	-0.382	-0.122	<.001	-0.325	-0.415	-0.235	<.001
Competence	0.698	0.578	0.819	<.001	0.516	0.429	0.604	<.001
Sociability	-0.152	-0.282	-0.022	0.023	-0.123	-0.203	-0.043	0.003
Morality	0.204	0.046	0.362	0.012	0.111	0.003	0.219	0.045
Part age	-0.247	-0.565	0.07	0.138	-0.054	-0.275	0.167	0.635
Part gender	-0.099	-0.418	0.22	0.548	0.152	-0.072	0.376	0.187
Part science engagement	0.026	-0.294	0.345	0.877	0.128	-0.084	0.34	0.239

Interesting Research								
Predictor	Study 1				Study 3			
	B	95%CI Low	95%CI High	*p*	B	95%CI Low	95%CI High	*p*
Age	0.047	-0.072	0.166	0.446	0.074	0.012	0.137	0.021
Gender	-0.115	-0.242	0.011	0.084	-0.051	-0.141	0.039	0.268
Ethnicity	-0.009	-0.079	0.06	0.792	0.032	-0.014	0.078	0.176
Discipline	0.013	-0.036	0.062	0.613	-0.013	-0.044	0.018	0.406
Attractiveness	0.374	0.233	0.516	<.001	0.213	0.142	0.284	<.001

Competence	0.136	0.022	0.251	0.026	0.200	0.122	0.277	<.001
Sociability	0.059	-0.109	0.226	0.496	0.049	-0.032	0.131	0.236
Morality	0.124	-0.007	0.255	0.068	0.132	0.039	0.225	0.006
Part age	0.030	-0.332	0.393	0.872	0.020	-0.226	0.265	0.876
Part gender	-0.169	-0.564	0.226	0.409	0.273	0.024	0.523	0.034
Part science engagement	0.382	-0.02	0.785	0.073	0.232	-0.017	0.482	0.071

Table 15. Coefficients, confidence intervals and p-values for the fixed effects of the predictors (age, gender, ethnicity, attractiveness, discipline, competence, sociability, morality, participant age, gender and science engagement) for the random intercepts model, when using mixed-effects data to predict 'looking like a good scientist' and 'looking likely to produce interesting research', for both Study 1 and 3.

Discussion

Study 3 provided additional evidence for the trends identified in the previous face-rating studies, confirming that a three-factor model of social judgement (competence, sociability and morality) is appropriate when investigating the first impressions the public forms of scientists. Moreover, it is reassuring to see that the same social dimensions predicted looking like a good scientist (low attractiveness and high competence) and looking likely to produce interesting research (high attractiveness and high competence), across a range of face stimuli and samples of participants, indicating the robustness of the effects.

Studies 1 and 3 pooled data

Given that both studies 1 and 3 tapped into the same concepts, and aimed to illustrate what first impressions people form of scientists from their facial appearance, one would expect to draw very similar conclusions from both studies. Thus, we conducted two mixed effects analyses on the pooled data from both studies, to investigate whether the effects of interest were replicated across the two studies. Firstly, a model predicting the good scientist or interesting research ratings (separately for each judgement) from the fixed effects of the

scientist's perceived age, gender, ethnicity, discipline, attractiveness, competence, sociability and morality, as well as the participant's age, gender and engagement with science. Additionally, the model included random intercepts for each participant (with uncorrelated random slopes for each by-participant effect of the scientist's perceived age, gender, ethnicity, discipline, attractiveness, competence, sociability and morality) and for each scientist face (with uncorrelated random slopes for each by-face effect of the participant's age, gender and engagement with science). The second mixed effects model contained the same fixed and random effects as the one described above, as well as fixed effects of study and by-study interactions with scientist's perceived age, gender, ethnicity, discipline, attractiveness, competence, sociability and morality, and participant's age, gender and engagement with science. Both models were fitted using *lme4* (Bates et al., 2015), while significance was assessed through both Satterthwaite approximated p-values and Wald confidence intervals (*lmerTest;* Kuznetsova et al., 2015). The model including the effect of study did not provide a significantly better fit than the model without an effect of study, for either judgements of good scientist (χ^2 (12) = 8.52, p = .744) or judgements of interesting research (χ^2 (12) = 11.54, p = .484). Both models suggested similar outcomes: scientists higher in perceived competence and morality, but lower in perceived physical attractiveness and sociability were perceived to look more like good scientists, while scientists rated higher on perceived competence, morality and physical attractiveness were thought to look more likely to produce interesting research (Table 16 and Figure 3).

Studies 1 and 3 Pooled Data – No Effect of Study

Predictor	Good Scientist				Interesting Research			
	B	95%CI Low	95%CI High	*p*	B	95%CI Low	95%CI High	*p*
Age	0.094	0.027	0.160	0.006	0.073	0.017	0.130	0.012
Gender	-0.004	-0.077	0.070	0.921	-0.063	-0.138	0.012	0.101
Ethnicity	0.056	0.005	0.107	0.034	0.026	-0.017	0.069	0.232
Discipline	0.029	-0.006	0.063	0.102	-0.007	-0.034	0.019	0.586
Attractiveness	-0.331	-0.414	-0.247	<.001	0.266	0.197	0.336	<.001
Competence	0.600	0.518	0.681	<.001	0.215	0.139	0.292	<.001
Sociability	-0.139	-0.206	-0.072	<.001	0.057	-0.018	0.132	0.140
Morality	0.167	0.073	0.262	0.001	0.149	0.058	0.240	0.002
Part age	-0.009	-0.193	0.175	0.927	-0.024	-0.235	0.186	0.820
Part gender	0.072	-0.112	0.257	0.444	0.185	-0.032	0.402	0.097
Part science engagement	0.100	-0.083	0.283	0.284	0.265	0.048	0.481	0.018

Studies 1 and 3 Pooled Data – Effect of Study included

Predictor	Good Scientist				Interesting Research			
	B	95%CI Low	95%CI High	*p*	B	95%CI Low	95%CI High	*p*
Age	0.115	0.040	0.191	0.003	0.061	-0.005	0.128	0.070
Gender	-0.022	-0.109	0.064	0.613	-0.084	-0.173	0.006	0.069
Ethnicity	0.056	0.002	0.110	0.044	0.015	-0.032	0.062	0.531
Discipline	0.031	-0.007	0.069	0.109	-0.0004	-0.031	0.030	0.979
Attractiveness	-0.297	-0.392	-0.202	<.001	0.285	0.204	0.365	<.001

Competence	0.592	0.503	0.682	<.001	0.177	0.090	0.263	<.001
Sociability	-0.135	-0.215	-0.056	0.001	0.051	-0.041	0.143	0.276
Morality	0.158	0.058	0.258	0.002	0.139	0.044	0.234	0.005
Part age	-0.258	-0.654	0.138	0.204	0.028	-0.620	0.677	0.932
Part gender	0.024	-0.218	0.266	0.846	0.048	-0.254	0.350	0.756
Part science engagement	0.087	-0.151	0.325	0.475	0.319	0.018	0.620	0.040
Study	0.290	-0.147	0.726	0.195	-0.195	-0.876	0.486	0.576
Study*Age	-0.054	-0.130	0.021	0.160	0.015	-0.051	0.081	0.650
Study*Gender	0.046	-0.041	0.132	0.301	0.032	-0.058	0.122	0.485
Study*Ethn	-0.009	-0.063	0.046	0.759	0.023	-0.024	0.069	0.347
Study*Att	-0.081	-0.176	0.014	0.098	-0.037	-0.118	0.043	0.365
Study*Disc	-0.008	-0.046	0.031	0.694	-0.013	-0.044	0.017	0.391
Study*Comp	0.044	-0.045	0.134	0.333	0.070	-0.016	0.156	0.114
Study*Soc	0.003	-0.077	0.083	0.945	0.001	-0.091	0.093	0.979
Study*Mor	-0.013	-0.112	0.087	0.803	0.034	-0.061	0.129	0.483
Study*P_Age	0.206	-0.190	0.602	0.310	0.00008	-0.648	0.649	0.999
Study*P_Gen	0.126	-0.116	0.367	0.310	0.204	-0.098	0.506	0.187
Study*P_Sci	0.029	-0.209	0.267	0.812	-0.095	-0.397	0.206	0.537

Table 16. Coefficients, confidence intervals and p-values for the fixed effects of the predictors (age, gender, ethnicity, attractiveness, discipline, competence, sociability, morality, participant age, gender, science engagement and study) as well as the interactions between study and all other fixed effects, for the random intercepts model, when using mixed-effects data to predict 'looking like a good scientist' and 'looking likely to produce interesting research', for the pooled data from Study 1 and 3.

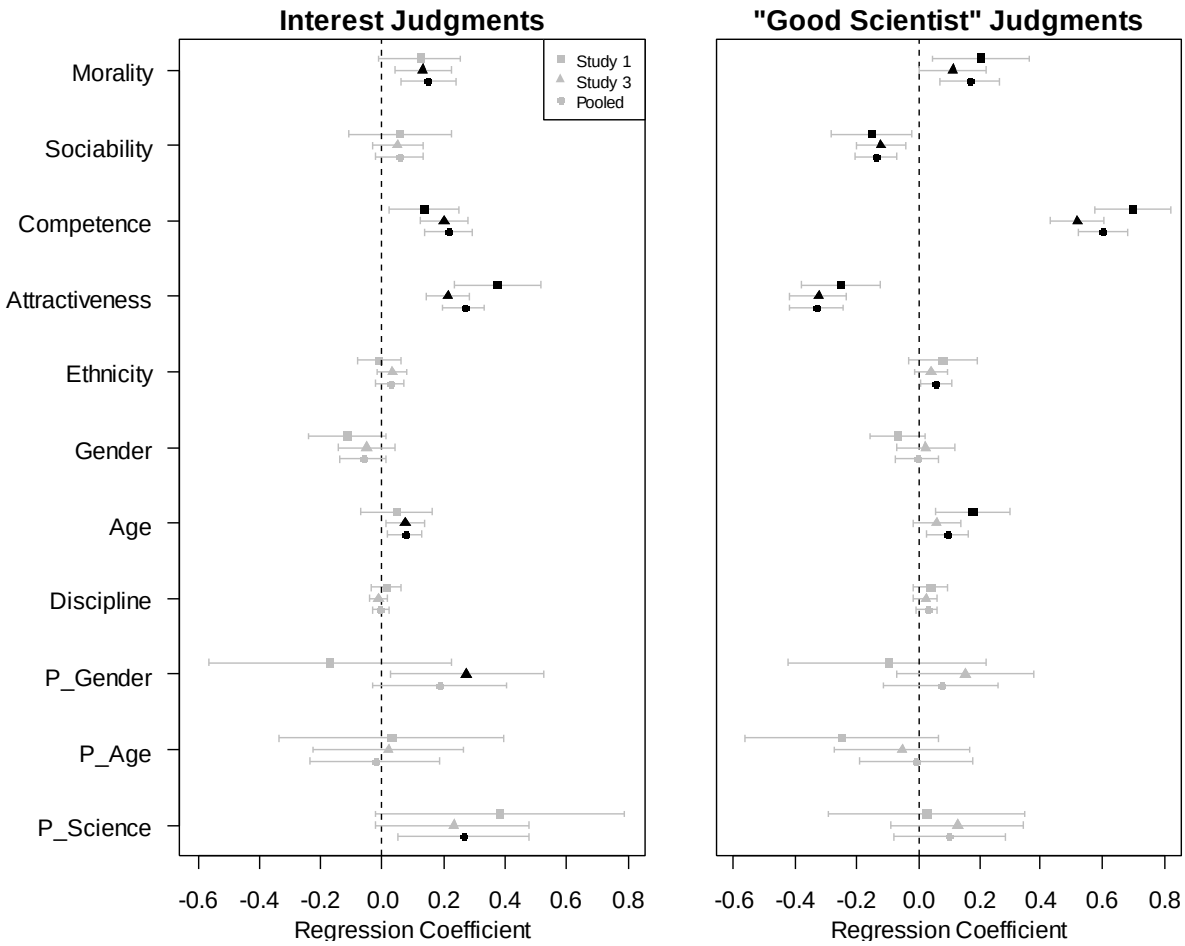


Figure 3. Regression coefficients and 95% Wald confidence intervals for the Study 1 and 3 data, and the pooled data. Coefficients with CIs that exclude zero are highlighted in black.

Furthermore, it was reassuring that no effect of study was found, and none of the interactions were significant, suggesting that none of the effects of interest were dependent on specific characteristics of the study. In turn, this increases the generalisability of our findings, which have been reproduced with both UK and US scientists, as well as both UK and US participants.

Chapter Summary

All in all, three face-rating studies were run, in an effort to better understand what first impressions people form of scientists, based on their facial appearance. The highest degree of agreement was between the two studies which used photos of real scientists (US scientists – Study 1, UK scientists – Study 3). Scientists who looked more competent and

moral, but less attractive and sociable were perceived to look more like "good" scientists (Table 17), while those who looked more competent and more attractive were perceived to be scientists who people would show interest in (Table 18). Study 2 provided evidence for the same direction of these trends, although not all the desired effects reached significance. Arguably this could be due to the fact that Study 2 was methodologically more different to the other two, since it used photos of people from a face database, as opposed to photos of real scientists. Additionally, the gender, age and race distribution of the stimuli in Study 2 were not representative of the population of scientists (e.g., more women and African American individuals); the participant samples were smaller as well, leading to more noise in the data. Alternatively, when looking at the mean ratings for the "good" scientist measure in Studies 1, 2 and 3, the difference between Study 2 (4.47) and Studies 1 and 3 (5.62 and 5.79, respectively) stands out the most. This trend speaks to the possibility that Study 2 represents a boundary condition to the effects observed in Studies 1 and 3; in other words, the majority of faces used in Study 2 did not resemble a stereotypical scientist. If this is the case, it would imply that some of the effects not replicated in Study 2 were localised to samples of scientists, and that participants were able to detect this in the faces presented.

Predictor	Study 1		Study 2		Study 3	
	B	*p*	B	*p*	B	*p*
Age	0.177	.007	-0.163	.185	0.059	.140
Gender	-0.068	.143	-0.249	.009	0.023	.633
Ethnicity	0.079	.179	-0.116	.034	0.040	.146
Discipline	0.039	.160	-	-	0.024	.283
Attractiveness	-0.252	<.001	-0.161	.112	-0.325	<.001
Competence	0.698	<.001	0.914	<.001	0.516	<.001
Sociability	-0.152	.023	-0.204	.085	-0.123	.003

Morality	0.204	.012	0.233	.119	0.111	.045
Part age	-0.247	.138	0.055	.799	-0.054	.635
Part gender	-0.099	.548	-0.243	.262	0.152	.187
Part science engagement	0.026	.877	0.417	.060	0.128	.239

Table 17. Coefficients and p-values for the fixed effects of the predictors (age, gender, ethnicity, discipline, attractiveness, competence, sociability, morality, participant age, gender and science engagement) on the "good scientist" measure, for the random intercepts and slopes model, compared across all three face-rating studies (1, 2 and 3).

Predictor	Study 1		Study 2		Study 3	
	B	*p*	B	*p*	B	*p*
Age	0.047	.446	0.198	.122	0.074	.021
Gender	-0.115	.084	-0.048	.492	-0.051	.268
Ethnicity	-0.009	.792	0.051	.213	0.032	.176
Discipline	0.013	.613	-	-	-0.013	.406
Attractiveness	0.374	<.001	0.199	.040	0.213	<.001
Competence	0.136	.026	0.197	.022	0.200	<.001
Sociability	0.059	.496	0.079	.263	0.049	.236
Morality	0.124	.068	0.003	.973	0.132	.006
Part age	0.030	.872	0.472	.068	0.020	.876
Part gender	-0.169	.409	0.279	.290	0.273	.034
Part science engagement	0.382	.073	0.408	.113	0.232	.071

Table 18. Coefficients and p-values for the fixed effects of the predictors (age, gender, ethnicity, discipline, attractiveness, competence, sociability, morality, participant age, gender and science engagement) on the "interesting research" measure, for the random intercepts and slopes model, compared across all three face-rating studies (1,2 and 3).

Summarizing the findings of the three face-rating studies, both interest in and beliefs about the quality of the work were related to the facial traits of the scientist: scientists who appear competent, moral and attractive are more likely to raise interest in their work, while scientists who appear competent and moral, but apparently unattractive and unsociable were perceived as doing higher-quality research. Please note that both Studies 1 and 3 used photos of real scientists collected from University websites, aiming to achieve a sample of ecologically valid stimuli. The ethical implications of this approach have been considered, and we have chosen not to reproduce any of the photos used in the thesis, as we do not have permission to do so.

CHAPTER 3: EFFECTS OF FACE-BASED FIRST IMPRESSIONS (LOOKING LIKE AN "INTERESTING" SCIENTIST) ON THE PUBLIC'S CHOICE OF SCIENTIFIC COMMUNICATIONS

Study 4: Is the public's choice of articles influenced by the appearance of the scientist?

Study 4 was designed as a validation study, aiming to expand the findings from Study 1 by investigating whether people's opinions of scientific research can be biased by the appearance of the researcher. We paired scientific article titles with photos of actual scientists (rated in Study 1) to examine whether people are more likely to select an article to read when associated with a face rated highly likely to produce interesting research. An initial pilot study was conducted to select a set of scientific article titles equally interesting, as described below. We expected participants to choose titles paired with "interesting" faces more often, when all the titles presented were equally interesting.

Pilot study

1. ***Pilot method***

    a. *Participants*

Participants were recruited using Amazon's Mechanical Turk. For the pilot study, the final sample consisted of 105 people (51 men and 54 women), with age ranging between 18 and 65 (M = 39.3, SD = 13.5), and 92% of participants were American (96% had English as their first language). Participants were paid for the completion of the questionnaire.

    b. *Stimuli and Materials*

For the pilot study, scientific article titles were selected from ScienceDaily.com - 30 articles were selected from the "Health and Medicine" category, and another 30 were selected from the "Physics" category, in the interest of rating the titles of the articles. The articles were selected by the researcher, aiming for a set of equally interesting titles, which did not mention the researchers conducting the study in third person (e.g., "Scientists discover..",

"Physicists show.."). None of the titles were edited or modified in any way; the final selection can be found in Appendix B.

c. Design and Procedure

Half the participants saw the biology titles, whereas the other half saw the physics titles, and rated them in terms of interest. The participants were presented with one article title at a time, and asked "How interested would you be in reading this article?". After making their response on a 10-point Likert scale (ranging from 0 - "Not at all interested" to 10 - "Extremely interested"), participants clicked on the ">>" button to progress to the next article title. Demographic information was also collected, and participants were debriefed on the last screen. The order in which the titles were presented was randomized for each participant, and participants' allocation to rating either biology or physics titles was counterbalanced.

For the biology titles, the response scale did not appear on screen for one of the titles (number 3, see Appendix B), so only data for 29 titles was collected for the biology category.

2. ***Pilot results***

a. Data preparation

A variance check was performed to ensure that participants had engaged with the task, and the results suggested that all participants did so.

b. Internal Reliability

To assess the reliability of the article ratings, two internal reliability measures were employed: a non-robust Cronbach's Alpha test and an inter-class correlation measure (ICC; Hallgren, 2012). This was performed separately for biology title ratings and physics title

ratings. Both types of scientific article titles had very good internal reliability on both measures, as illustrated in Table 19.

Internal Reliability Measure	Biology Titles	Physics Titles
Cronbach's Alpha	0.800	0.844
ICC	0.800	0.844

Table 19. Non-robust values of Cronbach's Alpha and inter-class correlations for biology and physics article ratings in the pilot of Study 4.

Main Study

1. Methods

    a. *Participants*

The minimum required sample size of 325 was based on a power analysis to achieve 95% power to detect a small-to-medium effect (w = 0.2) in a chi-square test looking at whether face type influences article choices.

For the main study, participants were recruited online, using Amazon's Mechanical Turk. Following the exclusion procedure, a final sample was comprised of 384 participants (218 men and 166 women). Approximately 6% of the sample was comprised of non-US citizens, with approximately 3% having a first language other than English. Participants' ages ranged between 18 and 82 (M = 32.7, SD = 10.7), and they were paid for their participation.

    b. *Stimuli and Materials*

The stimuli used in the main study were a selection of the titles used in the pilot study: 6 titles were selected from the Physics category, and 6 were selected from the Biology category. The chosen titles had similar ratings in terms of how interesting they seemed, as determined by the pilot study (the ratings were between 5.1 and 5.4 on a 10-point scale, see Table 20).

Biology Article Titles	Mean Rating (SD) out of 10
Opinions on vaccinations heavily influenced by online comments†	5.12 (3.20)
Confidence in government linked to willingness to vaccinate†	5.17 (3.29)
Texting may be more suitable than apps in treatment of mental illness†*	5.19 (2.94)
Cow immune system inspires potential new therapies†*	5.27 (2.73)
Reasons why winter gives flu a leg up could be key to prevention†*	5.35 (2.85)
Stress balls, DVDs and conversation ease pain, anxiety during surgery†*	5.37 (2.98)
Risk for autism increases for abandoned children placed in institutions	5.38 (3.04)
Elementary teachers' depression symptoms related to students' learning	5.52 (2.73)
Physics Article Titles	Mean Rating (SD)
Laser pulse turns glass into a metal: New effect could be used for ultra-fast logical switches†	5.13 (2.85)
Doing more with less: Steering a quantum path to improved internet security†	5.17 (2.96)
A 'Star Wars' laser bullet -- this is what it really looks like†	5.23 (3.09)
'Solid' light could compute previously unsolvable problems†	5.25 (2.99)
How to make mobile batteries last longer by controlling energy flows at	5.26 (3.04)

nano-level†	
Universe may face a darker future: Is dark matter being swallowed up by dark energy? †	5.32 (3.10)
Hunt for Big Bang particles offering clues to the origin of the universe	5.45 (3.10)
Electronics that need very little energy? Nanotechnology used to help cool electrons with no external sources	5.45 (3.00)

Table 20. Mean interest ratings and standard deviations for the article titles used in Studies 4, 5 and 6. All titles were used in Study 5; titles marked with a cross were used in Study 4, while those marked with an asterisk were used in Study 6.

The photos used in Study 4 were selected from the sample of photos rated in Study 1. The three highest-rated (means: 6.07, 6.04, 5.93 and standard deviations: 1.75, 1.60, 2.22, respectively) and three lowest-rated men (means: 3.67, 3.48, 2.96 and standard deviations: 2.09, 2.10, 1.74, respectively), as well as the three highest-rated (means: 6.67, 6.48, 6.26 and standard deviations: 1.73, 1.83, 1.58, respectively) and three lowest-rated (means: 4.33, 4.26, 4.19 and standard deviations: 1.69, 2.07, 2.45, respectively) women on the "How interested would you be in finding out more about this person's research?" dimension were selected, regardless of their discipline.

c. Design and Procedure

The study employed a 2x2 between-subjects design, with gender (male photos vs. female photos) and discipline (biology titles vs. physics titles) as factors. Thus, participants saw either male faces paired with biology titles, male faces paired with physics titles, female faces paired with biology titles or female faces paired with physics titles. The titles and photos were paired together using a Latin Square design, to ensure that each possible title-photo combination was represented, resulting in six possible combinations of faces and

titles. Each participant was presented with a single screen showing one of the six combinations of faces and titles, and was asked to select which article he/she would like to read. The allocation of participants to one of the four gender-discipline combinations and to one of the six article-faces combination was counterbalanced. Participants were led to believe that they would have to read the article of their choice, and to answer questions about it. After making their choice, the participants were debriefed and cleared of any deception, and had the opportunity to follow an external link to the article they had chosen, should they wish to read it.

Demographic information was also collected from the participants, alongside information regarding their engagement with science (see Appendix A).

2. Results

    a. *Science Engagement Questionnaire*

In order to confirm that the Science Engagement Questionnaire (see Appendix A) was indeed measuring a single concept of "Science Engagement" as found in Study 1, a Principal Components Analysis was conducted on the eight questions, to confirm its underlying structure. The PCA automatically extracted two factors with Eigenvalues above 1, which accounted for 73% of the total variance. The scree plot suggested that a one factor solution may be more appropriate, especially since the question loadings onto the two factors did not provide a satisfactory simple structure. Thus, a one-factor solution was imposed onto the data (accounting for 55% of the total variance), and the resulting analysis was satisfactory, as indicated by two measures of appropriateness of factor analysis: a Keiser-Meyer-Olkin Measure of Sampling Adequacy test (.872) and Bartlett's Test of Sphericity ($\chi^2(28) = 1694$. $p < .001$). The internal reliability of the one-factor scale was calculated, resulting in a Cronbach's $\alpha = .88$ for science engagement, suggesting very high internal reliability. Since the questions reliably measure a single concept, the composite measure

was computed by averaging the ratings across all eight items, and labelling the final variable Science Engagement.

b. *Chi-square analysis*

The primary research question was whether the proportion of people choosing an article associated with a photo rated highly on "looking likely to produce interesting research" was equal to the proportion of people choosing an article associated with a photo rated low on the same dimension. A chi-square test suggested that participants tended to choose the higher rated photos more often (~54% of participants chose an article associated with a higher rated photos, while ~46% of participants chose an article associated with a low rated photo), but not significantly so: $\chi 2(1) = 2.667$, $p = .103$.

c. *Logistic regression*

To investigate whether the tendency to choose highly rated photos over low rated photos was moderated by other variables, a logistic regression was conducted. The outcome for each participant (selecting an article associated with a high rated photo or one associated with a low rated photo) was predicted as a function of both participant variables (participants' gender, age and engagement with science, as measured by the questionnaire in Appendix A) and design variables (discipline of the article, and gender of the scientist in the photos). The overall model was not significant ($\chi 2(5) = 2.247$, $p = .814$), and neither were any of the individual predictors (as illustrated in Table 21). These results suggest that the participants did not choose high-rated faces more often than low-rated faces, and their choices were not moderated by any design or participant variables.

Predictor	B	*p*
Participant gender	-.193	.363
Participant age	.008	.417

Participant science engagement	-.044	.784
Article Discipline	-.141	.495
Photo Gender	-.109	.597

Table 21. Regression coefficients and levels of significance for all the participant-level predictors (gender, age, science engagement) and design-level predictors (article discipline and photo gender) entered in the logistic regression for Study 4.

Discussion

The results of Study 4 indicated that, although participants had a slight tendency to select titles associated with faces looking more likely to produce interesting research, this difference was not significant, nor influenced by design or participant variables. This outcome is not entirely surprising, considering that switching from a rating task to a single, forced choice task has led to a considerable decrease in experimental power. Furthermore, it is possible that the prospect of reading an article does not induce a focus on the appearance of the researcher. A different medium (e.g., watching a video or a podcast of the scientist talking about their research) may induce a higher focus on appearance. This hypothesis was investigated in Study 5.

Study 5: Is the public's choice of communication influenced by the appearance of the scientist? Does this differ between articles and videos?

The fifth study was designed as a replication of Study 4, aimed to increase experimental power by increasing the number of responses collected per participant to four. Additionally, we introduced an element of medium: participants either chose an article to read or a video to watch. We explored whether choosing a video to watch would increase the focus on the scientist's facial appearance, given the prospect of having to watch the scientist talk about their research, and therefore would yield a larger effect of the type of face the communication was paired with. With this set up, participants are more likely to rely on

superficial cues such as physical appearance when making judgements about the articles/videos presented (e.g., Chaiken & Trope, 1999).

Method

1. *Participants*

The required sample size was based on a power analysis to achieve 80% power to detect a small effect (w = 0.1) in a chi-square test looking at whether face type influences article choices. The analysis suggested a target sample size of a minimum 785 participants. Participants were recruited using Amazon's Mechanical Turk, and no participants were excluded.

The final sample consisted of 849 participants (526 men and 323 women), with age ranging between 18 and 73 (M = 32.4, SD = 10.6). Approximately 93% of the participants reported having English as their first language, while about 82% of the sample stated they were US citizens. Out of the sample, 427 were assigned to the Text condition, and 422 to the Video condition.

2. *Stimuli and Materials*

Considering the aim of this study was to replicate and expand the findings from Study 4, the same 6 titles from the Physics category and 6 from the Biology category were used. In addition to these 12 titles, a further 4 were selected, 2 from the Biology category and 2 from the Physics category, to allow for a full counterbalancing of titles across conditions (Table 20).

Similarly, the same 12 photos used in Study 4 were used, with an additional 4 photos: the fourth highest/lowest rated on the same dimension, for both men and women.

3. Design and Procedure

The study consisted of a 2x2x2 mixed between and within-subject design, with task (choose an article to read vs. choose a video to watch) as a between subjects factor, and discipline (biology titles vs. physics titles) and gender (male photos vs. female photos) as within-subjects factors. All participants saw the four combinations of discipline and gender (male faces paired with biology titles, male faces paired with physics titles, female faces paired with biology titles or female faces paired with physics titles); half of the participants were told they would choose articles to read, while the other half were told they would choose videos to watch. Participants saw each discipline-gender combination at a time, and each time they had to choose from four face-title combinations, making four choices in total (one for each discipline-gender combination). Each participant saw all the 16 faces and 16 titles, without any repetitions. Twenty-four versions of the task were created (the same 24 versions were used for the two different tasks), ensuring that each face occurred equally often in Biology and Physics, that each article occurred equally often with each combination of gender and face-type[2], and that neither photos nor titles were repeated for any participant. This was achieved using a combination of counterbalancing and Latin-square designs. Additionally, the order in which the four blocks (i.e., discipline-gender combinations) were presented was randomized individually for each participant. The order of the four options (i.e., face-title combinations) on the page was also randomized separately for each block. The allocation of participants to one of the two mediums was counterbalanced. The study resulted in four data-points per participant: their four choices of articles to read, or videos to watch, respectively. Please note that we were interested solely in their *choice* of scientific communication, as opposed to the participants' *opinion* of the research quality.

[2] Face-type here refers to photos rated high versus photos rated low on looking likely to produce interesting research

To increase the ecological validity of the task, participants were informed that they would be presented with audio information during parts of the study, and were asked to perform an audio check. The check consisted of listening to the word "Welcome" and selecting the word they listened to.

We collected demographic information and participants' interest in and engagement with science (See Appendix A). Participants were cleared of any deception (i.e., that there were no videos to watch, only articles, and the people associated with the titles had no connection to the scientific research presented), and had the opportunity to read the articles they had selected by following hyperlinks within the feedback letter.

Results

1. *Mixed effects logistic regression*

A mixed effects logistic regression was used to verify whether communications paired with high-rated faces on the "interesting research" measure were chosen significantly more often than communications paired with low-rated faces. We also investigated whether the effect differed between levels of the factors (i.e., male vs. female, biology vs. physics), and whether participant-level variables predicted the effect. We used the R *glmer* function with the "*bobyqa*" optimizer due to the rigidity of the latest *lme4* package (Bates et al., 2015).

The outcome for each choice (selecting a communication associated with a high rated photo, or one associated with a low rated photo) was predicted as a function of the task (choosing and article or a video), discipline of the communication (biology or physics), gender of the scientist (male or female), as well as the participant's gender, age, science engagement (as measured by the questionnaire in Appendix A), and all the possible 2-way and 3-way interactions between factors. Given the structure of the data (four scores were collected per participant), the model also included a random intercept for participant, and random slopes for gender, discipline and their interaction. The significance of the model

estimates was assessed with both Satterthwaite-approximated p-values and Wald confidence intervals (Table 22, Figure 5). The results suggested that, overall, the odds of choosing a face rated as "high" were higher than those of choosing a face rated as "low". High faces were chosen over 50% (~54% for articles and ~60% for videos) of the time across all combinations of gender and discipline, and this pattern was similar for both articles and videos (Figure 4 – error bars were calculated using Morey's method of calculating confidence intervals for within-subjects conditions; Morey, 2008).

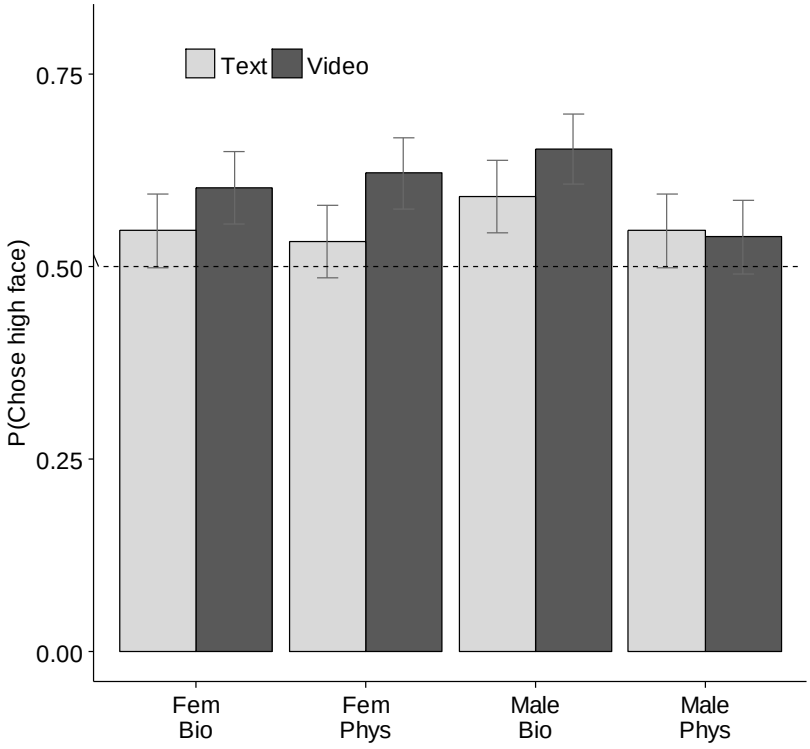


Figure 4. Proportion of "high" faces (on the interesting research dimension) chosen and confidence intervals, across all combinations of gender and discipline, for both article choices and video choices in Study 5.

Additionally, the odds of choosing a "high" face were higher for people in the video condition, and when people chose from the biology category. The odds of choosing a "high" face were also higher for women and younger participants, as illustrated in Table 22.

Predictor	B	95% CI Low	95% CI High	*p*
Intercept	0.333	0.257	0.408	<.001
Task	0.104	0.030	0.178	.006
Discipline	-0.096	-0.178	-0.013	.023
Gender	-0.017	-0.098	0.064	.682
Task * Disc.	-0.024	-0.106	0.059	.574
Task * Gender	0.056	-0.025	0.136	.178
Disc. * Gender	0.118	0.025	0.212	.013
Task * Disc. * Gender	0.075	-0.018	0.168	.116
Part age	-0.089	-0.164	-0.014	.020
Part gender	-0.134	-0.210	-0.058	<.001
Part science engagement	-0.028	-0.104	-0.048	.467

Table 22. Coefficients, confidence interval limits and p-values for the fixed effects of the factors (task, discipline, gender and their interactions), as well as participant level variables (participant age, gender and science engagement) for predicting the odds of choosing a face rated as "high" in Study 5.

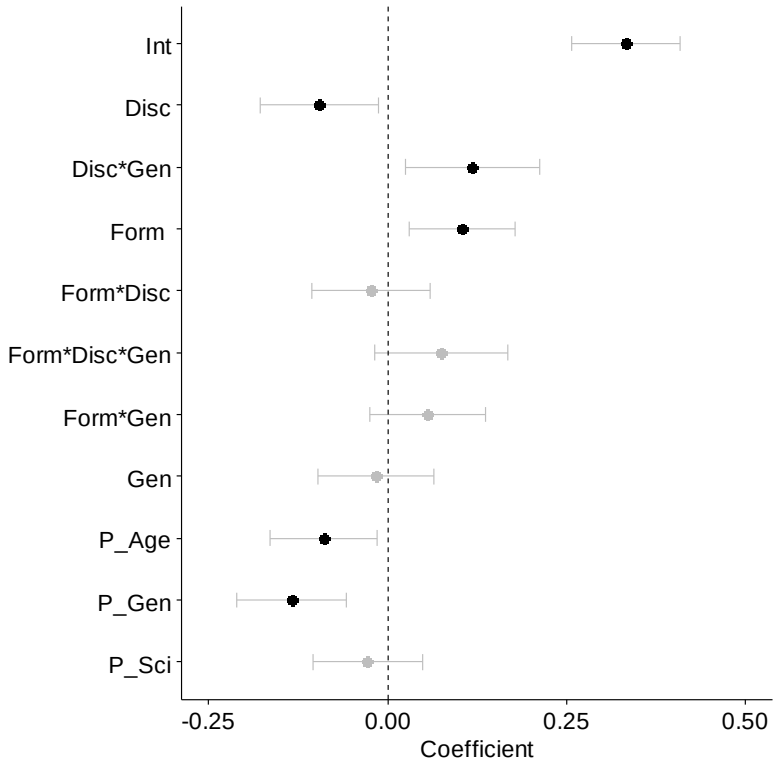
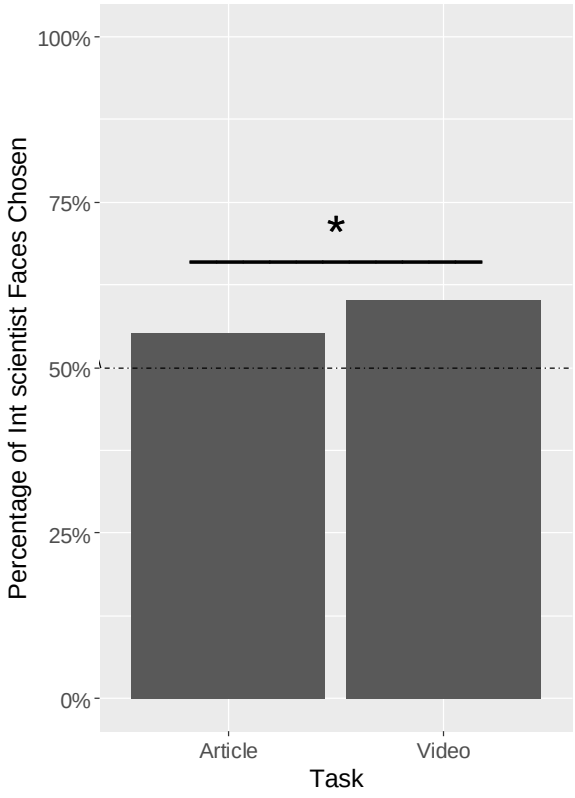


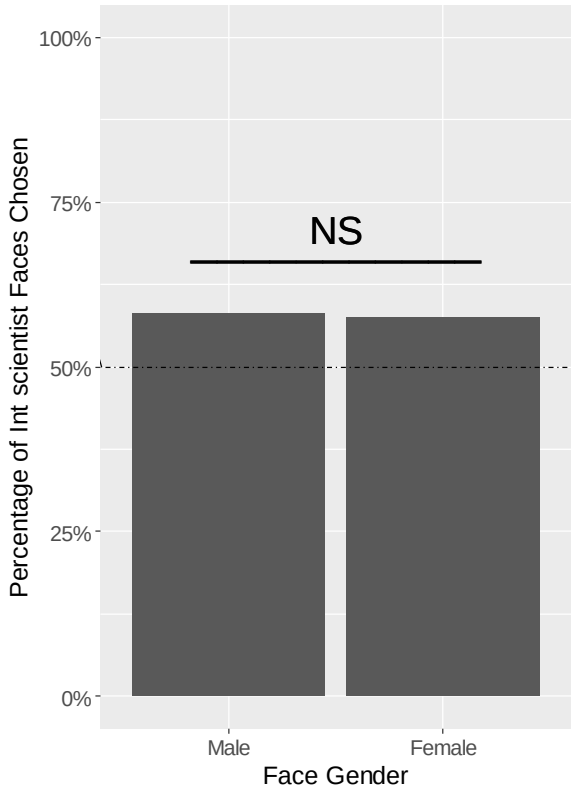
Figure 5. Regression coefficients and 95% Wald confidence intervals for the random intercept and random slopes model in Table 22. Coefficients with CIs that exclude zero are highlighted in black. "Form" here refers to the format of the task (article or video).

2. Simple Main Effects and Interactions

The full model fitted above suggests that, overall, the odds of choosing a face rated as 'high' were higher than the odds of choosing a face rated as 'low'. However, this analysis does not indicate whether this is the case for each level of each factor (Male faces, Female faces, Biology titles, Physics titles, choosing Articles or choosing Videos). The main effects (Table 22) indicate that the difference between 'low' face and 'high' face choices depends on whether the participant was in the Article or Video condition, and whether they were picking Biology or Physics articles, but it is unclear why this is the case. The difference could be due to (a) a tendency to choose 'high' faces in both tasks/for both disciplines, but the tendency is more pronounced for one level than the other, (b) an effect of facetype for one

task/discipline and not the other, or (c) opposite effects of facetype on the two tasks/disciplines. Judging from the response probabilities, option (c) is not a possibility: 'high' faces were chosen over 50% of the time for all simple main effects (see Figure 6).


100%
75%
50%
25%
0%
Percentage of Int scientist Faces Chosen
*
Article
Video
Task


100%
75%
50%
25%
0%
Percentage of Int scientist Faces Chosen
NS
Male
Female
Face Gender

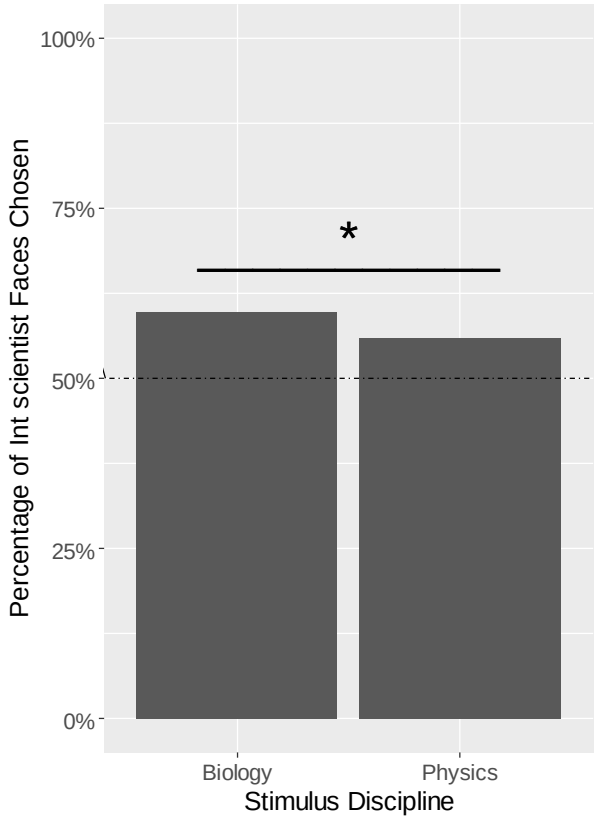


Figure 6. Simple main effects of Task (article; video), Gender (male faces; female faces) and Discipline (biology titles, physics titles), expressed relative to the percentage of "Interesting Scientists" faces chosen in Study 5.

The simple main effects of the three factors were investigated using dummy coding on the full model logistic regression (predictors, control variables, random intercept for participant and random slopes for gender, discipline and their interaction), to highlight the effect of interest. The simple main effects of Task (article, Intercept Estimate = 0.383, $p < .001$; video, Intercept Estimate = 0.644, $p < .001$), Gender (male, Intercept Estimate = 0.513, $p < .001$; female, Intercept Estimate = 0.308, $p < .001$) and Discipline (biology, Intercept Estimate = 0.513, $p < .001$; physics, Intercept Estimate = 0.173, $p = .015$) were all significant, suggesting that the differences inferred from the main effects are due to (a) a tendency to choose 'high' faces in both tasks/for both disciplines, but the tendency is more pronounced for one level than the other. Thus, 'high' faces were chosen more often in each of the six levels of the predictors.

The interaction between gender and discipline was also analysed within the context of the dummy coded logistic regressions using the full model, by investigating the effect of one variable at each level of the other. The results suggest that there is a significant difference between the proportion of 'high' faces chosen between biology titles and physics titles for male faces (Discipline Estimate at Male = -0.338, $p < .001$), but not for female faces (Discipline Estimate at Female = 0.014, $p = .893$). Similarly, there is a significant difference between the proportion of 'high' faces chosen between male faces and female faces for biology titles (Gender Estimate at Biology = -0.204, $p = .043$) but not for physics titles (Gender Estimate at Physics = 0.146, $p = .144$). These trends are reflected in the percentage of 'high' faces chosen in each cell of the design: "interesting" scientists were chosen 62% of the time for male biology, 54% for male physics, 57% for female biology and 58% for female physics.

Discussion

Study 5 found that participants chose scientific communications associated with scientists rated higher on "looking likely to produce interesting research" more often than communications associated with low rated scientists. This was particularly the case when participants believed they would be watching a video of the scientist they had selected. The proportion of articles associated with "high" faces chosen in Study 4 (~54%) was very similar to the proportion of articles associated with "high" faces in the current study (~ 54% as well), suggesting the failure to detect an effect in Study 4 was due to a lack of power.

Study 6: Is the public's choice of videos influenced by the perceived attractiveness and competence of the scientist?

Our previous studies have shown that scientists who are perceived to be more facially competent and more physically attractive are also perceived to look more likely to produce interesting research (i.e., someone whose research people want to find out more about).

Combining these findings with the ones from Study 5 (which showed that people are more likely to want to find out more about research associated with scientists high on interest ratings), in Study 6 we investigated whether a scientist's perceived attractiveness and competence predict how much interest the public shows in the scientist's research. In Study 6 we will focus on male scientists, associated with biology titles; this decision was informed by the interaction effect found in Study 5, where the strongest effect was observed in the male-biology condition ("interesting" scientists were chosen 62% of the time in this condition). More specifically, we expected research associated with highly competent and attractive scientists to receive more interest from the public. Study 6 was pre-registered on the Open Science Framework (osf.io/ev794).

Method

1. *Participants*

A power analysis conducted using GPower revealed that a minimum of 330 participants were needed for 95% power to detect a small-to-medium effect size (d=0.2, based on the modest effect found in Study 5) in a within-subjects t-test for an effect of face type, with a standard alpha criterion of .05. Since the effect size was roughly estimated from our previous studies, we aimed for a minimum sample of 400 participants. Participants were recruited using Amazon's MTurk, and two participants were excluded for reporting technical issues with the survey (i.e., the photos not loading properly).

The final sample was comprised of 408 participants (192 men and 216 women), with age ranging from 18 to 74 (M = 35.9, SD = 11.1). Approximately 88% of participants reported being US citizens, or of American descent, while 98% reported having English as their first language; all participants were paid at a standard rate for a 3-minute study.

2. *Stimuli and Materials*

The scientific titles used were four of the six biology titles in Study 5, matched in terms of interest ratings (Table 20) –the two titles with the most extreme scores were removed. The scientist photos were selected from the photos used in Study 3 (UK scientists), which had been rated on a number of social dimensions. Two photos were chosen for each of the following combinations of the variables of interest (attractiveness and competence, where high/low represent the top/bottom rated 12.5% of photos): high competence high attractiveness (HCHA), high competence low attractiveness (HCLA), low competence high attractiveness (LCHA), low competence low attractiveness (LCLA). Due to the gender imbalance in the sample of photos we used, only photos of males were selected, resulting in a total of 8 photos (see Table 23 for exact ratings).

	Low Competence		High Competence	
	Low Attractiveness	High Attractiveness	Low Attractiveness	High Attractiveness
Attractiveness	2.65 (1.31)	5.60 (2.15)	2.81 (1.30)	5.12 (1.72)
Competence	4.62 (2.03)	5.02 (1.75)	6.65 (1.75)	6.69 (1.56)
Interest	4.23 (2.41)	5.05 (2.36)	5.55 (2.51)	5.71 (2.37)
"Good Scientist"	4.96 (2.55)	4.34 (2.28)	7.16 (1.92)	6.06 (1.89)
Age	42.38 (7.92)	26.07 (4.27)	52.62 (7.68)	42.02 (6.52)
Sociability	5.80 (1.88)	4.61 (1.82)	5.64 (1.78)	4.91 (1.85)
Morality	5.16 (2.09)	5.23 (2.04)	6.14 (1.83)	5.74 (1.93)
Warmth	5.48 (1.99)	4.92 (1.93)	5.89 (1.80)	5.32 (1.89)

Table 23. Mean ratings (SD) for the face stimuli used in Studies 6 and 9.

3. Design and Procedure

Study 6 had a 2x2 fully within-subjects design, with attractiveness (high or low) and competence (high or low) as factors. The biology titles were paired with the cells of the

design using a 4x4 Latin Square design, creating four versions of the task; the task comprised of four trials, representing a photo of the scientist paired with a biology title. Each participant was randomly allocated to one of the four versions. For each cell of the design, the survey in Qualtrics was designed to randomly select one of the two available photos to be displayed alongside the title. The order in which the trials were presented was randomised for each participant. Participants were asked to imagine they were browsing a website hosting videos of scientists discussing their research. For each "video" (represented by the biology title and scientist photo combination), participants had to think about how interesting and enjoyable or boring and unenjoyable it would be to watch each scientist talk about their research, and then rate how likely they would be to choose to watch each video. The ratings were done a 7-point Likert scale, from 1 -Not at all likely to 7 – Extremely likely. The participants were debriefed and cleared of any deception, and had the opportunity to follow external links to the original articles on which the stimuli were based, should they wish to read them.

Demographic information was also collected, alongside information regarding their engagement with science (see Appendix A).

Results

1. *Mixed effects linear regression*

The main question this study aimed to answer is whether a scientist's perceived facial competence and physical attractiveness would predict how much interest the public would show in research associated with the scientist. We predicted that participants would show more interest in (i.e., would choose to find out more about) research associated with scientists rated high both on competence and attractiveness, given the previously shown connection between these dimensions and looking likely to produce interesting research (Figure 7).

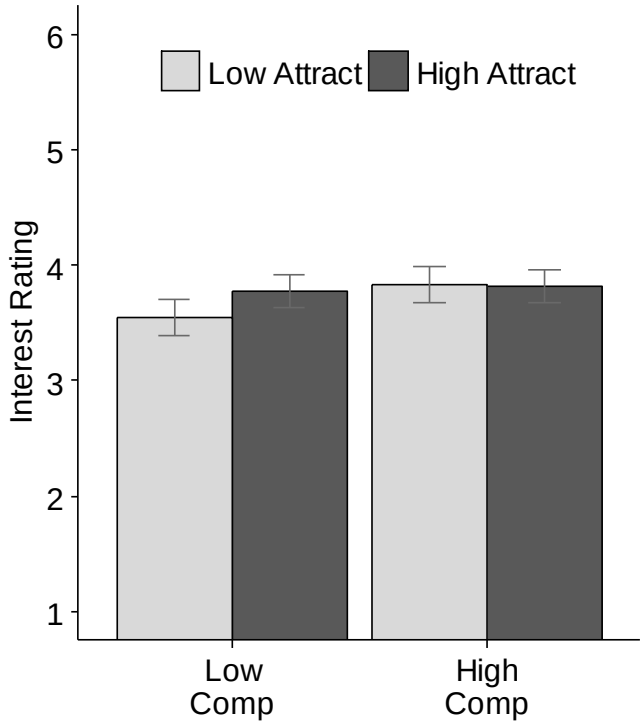


Figure 7. Mean interest ratings and error bars (Morey, 2008) for research associated with scientists that were either high or low on perceived competence and physical attractiveness in Study 6.

The data was analysed using a mixed effects linear model, fitted with the lme4 R package (Bates et al., 2015). Following our proposed analysis, the interest rating was regressed onto the perceived competence of the scientist (high or low), their physical attractiveness (high or low), the interaction between them, the gender, age, and science engagement of the participant (i.e., participant-level variables), as well as the interactions between competence and participant-level variables, and attractiveness and participant-level variables. Given the structure of the data, (four responses per participant), the model included a random intercept for participant, and random, uncorrelated slopes for competence, attractiveness and their interaction. Significance was evaluated using p-values produced using Satterthwaite approximations and 95% Wald confidence intervals (*lmerTest* R package). The categorical predictors were coded as 0 and 1, and scaled.

The analysis suggested that people showed more interest in research associated with competent-looking scientists. Although the trend for attractiveness was in the predicted direction (i.e., more interest for research associated with attractive scientists), the effect was not significant, and neither was the interaction between competence and attractiveness. Older participants and those who were more engaged with science were more likely to show interest in scientific research, regardless of the appearance of the scientists it was associated with (Table 24, Figure 8).

Predictor	B	95% CI Low	95% CI High	*p*
Intercept	3.743	3.629	3.858	<.001
Competence	0.083	0.007	0.158	.032
Attractiveness	0.052	-0.027	0.131	.196
Comp. * Att.	-0.059	-0.129	0.010	.093
Part gender	0.104	-0.014	0.223	.084
Part age	0.124	0.007	0.240	.039
Part science engagement	0.380	0.261	0.499	<.001
Comp. * Part gender	-0.051	-0.129	0.026	.196
Comp. * Part age	-0.023	-0.099	0.054	.564
Comp. * Part science engagement	0.029	-0.049	0.107	.471
Att. * Part gender	0.009	-0.072	0.091	.821
Att. * Part age	0.060	-0.021	0.140	.148
Att. * Part science engagement	-0.038	-0.120	0.043	.357

Table 24. Coefficients, 95% Wald intervals and p-values for the fixed effects of the factors (competence, attractiveness and their interaction), participant-level variables (gender, age

and science engagement), as well as 2-way interactions between the factors and participant-level variables, for predicting interest in research in Study 6.

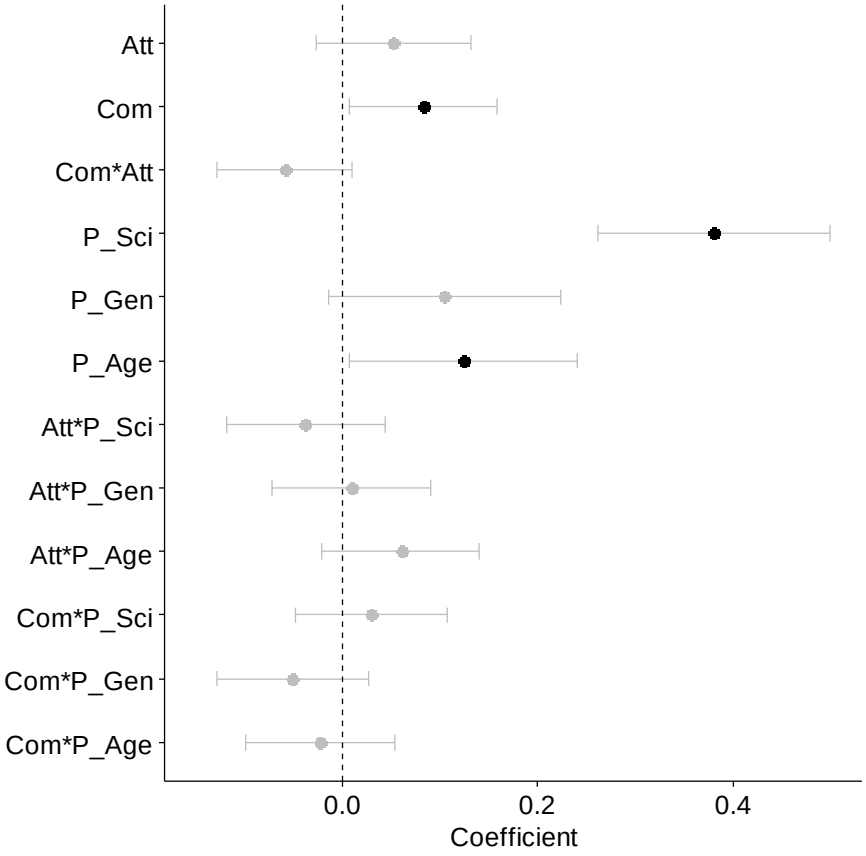


Figure 8. Regression coefficients and 95% Wald confidence intervals for the random intercept and random slopes model in Table 24. Coefficients with CIs that exclude zero are highlighted in black.

Discussion

To conclude, Study 6 found that participants were more likely to show interest in research associated with competent-looking scientists. Attractiveness of the scientists did not significantly affect interest ratings. Due to our ecological sample of photos, the attractiveness manipulation was weaker than the competence manipulation. The mean interest rating for the low-attractiveness faces was 4.89 while the mean interest rating for the high-attractiveness faces was 5.38, resulting in a difference of 0.49. However, the mean interest rating for the low-competence faces was 4.64, compared to 5.63 for the high-competence faces, resulting in a difference of 0.99. Since the interest ratings previously

predicted the choice of communication, it is possible that the low- vs. high-attractiveness photos did not differ sufficiently to generate an effect.

Chapter Summary

Taken together, the three studies that investigated what influences the public's choice when it comes to scientific news stories showed a similar trend: the facial appearance of the scientist associated with the research can influence the choices people make when selecting what scientific communications to find more about. People preferred to choose communications that were paired with scientists who looked more likely to produce interesting research (Study 5), and also with scientists who looked more competent (Study 6). While this was the case for both text- and video-based communications, the effects were stronger when participants believed they would be watching the videos they chose. Although at first glance a 60% preference for scientists looking more likely to produce interesting research might seem small, over a larger scale (such as TED talks and YouTube videos) it would translate into hundreds of thousands of extra views or shares on websites that host scientific material. Thus, face-based judgements appear to play an important role in modulating the public's engagement with scientific research, especially when shared on large-scale platforms.

CHAPTER 4: EFFECTS OF FACE-BASED FIRST IMPRESSIONS (LOOKING LIKE A "GOOD" SCIENTIST) ON THE PUBLIC'S OPINION ON SCIENTIFIC COMMUNICATIONS

Study 7: Are "good" scientists perceived to experience more positive scientific outcomes?

Study 7 was designed to investigate the relevance of first impressions based on facial appearance for the public's behaviour. In particular, Study 7 verified whether looking like a "good" scientist translates into "acting" like a good scientist as well. Therefore, we asked participants to choose which scientist is most likely to find himself/herself in one of the scenarios presented, expecting good scientists to be associated with positive research scenarios more often. Study 7 will also help validate our "good scientist" measure from Study 1, similarly to how Study 4 provided a validation for the "interesting research" measure.

Method

1. *Participants*

Participants were recruited online via Amazon's MTurk; one participant was removed after reporting technical problems (i.e., photos not loading).

The minimum required sample size of 199 was based on obtaining 80% power to detect a small-to-medium effect size (d = 0.2) in a within-subjects t-test for comparing two face types. For this study, the final sample comprised of 222 people (128 men, 94 women). Age ranged between 19 and 68 years old (M = 33.7, SD = 10.5). Approximately 93% of the participants reported being American citizens, while 97% had English as their first language.

2. *Stimuli and Materials*

The four highest and four lowest rated photos on "How much does this person look like a good scientist?" from Study 1 were used, for both men and women (16 photos in total, 8 from each gender, 4 high and 4 low). The lowest rated male photo had to be replaced with the fifth lowest rated photo, due to the person in the photo wearing a hat, which would

have been inconsistent with our other stimuli. Four different scenarios were also created for this study: two positive ("Which one of these scientists is most likely to have won a prize for their research?" and "Which one of these scientists is most likely to have recently published a paper in a renowned scientific journal?") and two negative ("Which one of these scientists is most likely to have been accused of plagiarism?" and "Which one of these scientists is most likely to have fabricated their results in order to publish a paper?").

3. *Design and Procedure*

Study 7 had a 2x2 within-subjects design, with photo gender (male photos vs. female photos) and scenario valence (positive scenarios vs. negative scenarios) as factors. All participants saw the four combinations of gender and scenario: male faces paired with a positive scenario, female faces paired with a positive scenario, male faces paired with a negative scenario and female faces paired with a negative scenario. Each combination was presented once at a time, and each time participants had four photos to choose from (two high and two low, all male or all female depending on the gender x scenario combination), resulting in four data points collected from each participant. To ensure that each face was presented equally often in positive and negative scenarios and that each face gender was equally paired with a one of the four scenarios, 12 versions of the task were created, by using and combining two separate counterbalancing designs. The order in which the four gender-by-scenario blocks were presented was randomized separately for each participant, while the order of the four faces on the screen was randomized for each block. Demographic information, as well as information about the participants' interest and engagement with science (see Appendix A) was also collected.

Results

1. *Mixed effects logistic regression*

We conducted a mixed effects logistic regression, investigating whether the choice of "good" faces versus "bad" faces depended on the scenario presented or on the gender of the face, and whether this was influenced by other participant-level variables. The outcome for each choice (choosing a "good" scientist or a "bad" scientist) was regressed onto the type of scenario, the gender of the scientist, their interactions, as well as the participants' gender, age and science engagement. Due to the structure of the data (four choices collected from each participant), we initially included a random intercept for participant, and random uncorrelated slopes for gender, scenario, and their interaction; however, due to the small number of data points, the model did not converge. Therefore, we ran a random intercept only model with the same predictors, which revealed that the random effects did not contribute anything to the model – the estimates and p-values were no different from a standard logistic regression with no random effects included. Hence, from here onwards we will report the results of the standard logistic regression. The analyses revealed that good scientists were chosen significantly more often in positive scenarios: across all participants and choices, an overwhelming 85% of the choices were congruent (i.e., "good" scientists selected in positive scenarios, and "bad" scientists selected in negative scenarios; Figure 9).

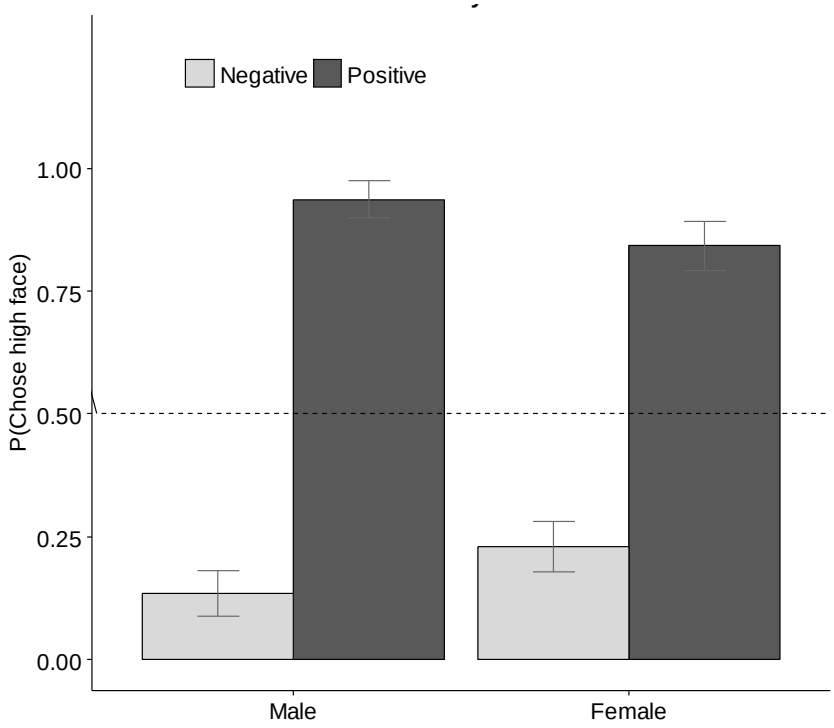


Figure 9. Proportion of Good Scientist choices and confidence intervals (Morey, 2008), for male and female scientists, presented in either positive or negative scenarios in Study 7.

However, high rated faces were chosen more often for women in negative scenarios, while high rated faces were chosen more often for men in positive scenarios (see Figure 10, Table 25).

Predictor	B	95% CI Low	95% CI High	*p*
Intercept	-1.442	-1.793	-1.109	<.001
Gender	-0.188	-0.606	0.216	.367
Scenario	3.728	3.330	4.152	<.001
Gender * Scenario	-0.418	-0.628	-0.217	<.001
Part Gender	-0.056	-0.249	0.136	.568
Part Age	-0.074	-0.267	0.117	.448
Part science engagement	0.027	-0.164	0.218	.784

Table 25. Coefficients, 95% Wald confidence intervals and p-values for the fixed effects of scenario, gender and their interaction, as well as participant level variables (participant age, gender and science engagement) for predicting the odds of choosing a face rated as "high/good" in Study 7.

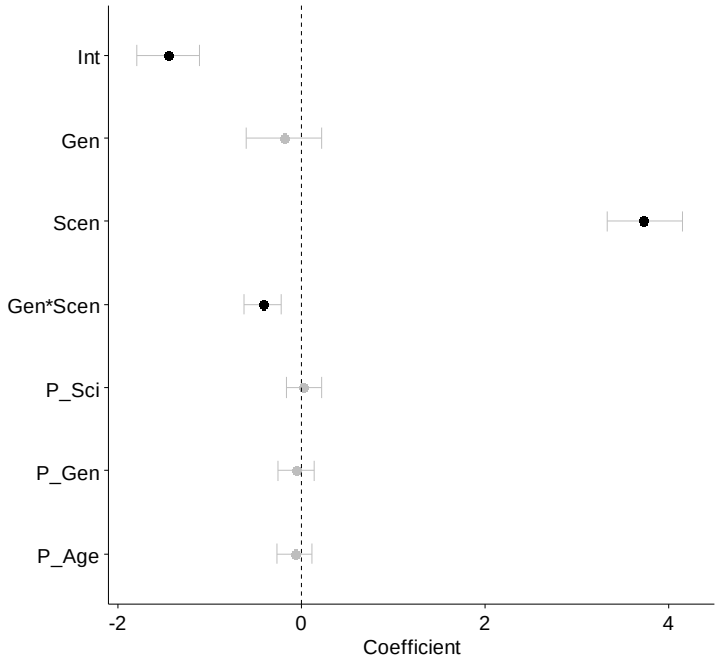


Figure 10. Regression coefficients and 95% Wald confidence intervals for the random intercept and random slopes model in Table 25. Coefficients with CIs that exclude zero are highlighted in black.

Discussion

The results of Study 7 illustrated that looking like a "good" scientists translates into being expected to behave like a good scientist as well. "Good" scientists were believed to have more positive academic experiences (such as publishing and winning prizes), while "bad" scientists were believed to have more negative academic experiences, such as plagiarising and fabricating data. The results also suggest that the "good" scientist measure we have developed and employed is indeed reflected in people's perception and expectation of scientists. . If people's expectations of scientists are influenced by their appearance, it is possible that their perception of scientists' work would also be influenced by the scientists' appearance; this will be investigated in Study 8.

Study 8: Is the public's opinion of scientific communications influenced by the appearance of the scientist?

Study 8 extended Study 7 by examining whether first impressions based on the facial appearance of scientists influenced people's perception of the scientists' work. Previous research has shown that pairing scientific articles with either male or female author names led to a bias in perceived quality: male-authored publications were rated as higher in scientific quality (Knobloch-Westerwick et al., 2013). If author gender alone can have such an effect on perceived quality, we expect facial appearance of the author to influence perceptions of the scientists' work. This was achieved by pairing photos of "good" or "bad" scientists with scientific articles matched in terms of their quality. A set of scientific articles similar in quality was selected following a pilot study, as described below. If looking like a "good" scientist translates into conducting good research, we expect articles paired with "good" scientists to be perceived as higher in quality, than articles paired with "bad" scientists.

Pilot study

1. Pilot method

a. Participants

For the pilot study, participants were recruited on-line, via Amazon's Mechanical Turk. Following exclusion, the final sample contained 128 people (68 men, 60 women). Participants' age ranged between 20 and 70 (M = 34.5, SD = 10.9), and 93% were American nationals (97% had English as their first language). Participants were paid for the completion of the questionnaire.

b. Stimuli and Materials

For the pilot study, a total of 20 scientific articles were selected from scientific news websites (such as newser.com) and re-written or modified in a more user-friendly manner. Ten Biology and ten Physics articles were re-written in first person, to simulate a "scientist profile" type of magazine section. The final selection can be found in Appendix C.

c. Design and Procedure

Participants saw half of the biology articles, and half of the physics articles, resulting in a total of 10 articles per person. The participants were presented with one article at a time, and asked the following questions: "How valid were this scientist's conclusions?", "How important are this scientist's findings?", "How good was the overall quality of this scientist's research?", "How rigorous was this scientist's research?", "How likely are you to describe this scientist's research to another person (e.g., a friend, colleague or family member)?", "How well did you understand this scientist's description of their research?" and "Have you heard/read about this research before? (Not just this general topic, but this specific piece of research.)"). All responses were made on 7-point Likert-type scale (ranging from 1-Not at all to 7 –Extremely), except for the last question which had a Yes/No answer. The five biology

articles and five physics articles were randomly chosen for each participant, and the order in which the articles were presented (i.e. biology first or physics first) was also randomly determined. The order of the evaluation questions was randomised individually for each participant. Demographic information was also collected, and participants were debriefed on the last screen.

2. Pilot results

a. Data preparation

Mean judgement ratings for each biology and physics article and for each question were computed, averaging across all the participants. Ratings where the participant mentioned having read about the research before were excluded from any future analyses.

b. Internal Reliability

To assess the reliability of the article ratings, a Cronbach's Alpha test was performed individually for each article, on the 5 items of interest (not including the recognition check question and the ease of comprehension question).The ratings of the items for each article had very good internal reliability, as illustrated in Table 26. Additionally, correlations between the evaluation items were computed, for both the aggregated data (using the means for each article, collapsed across participants; see Table 27, top half) and the disaggregated data (ignoring clustering by article and participant; see Table 27, bottom half). Although the p-values provided by the disaggregated data are essentially meaningless due to the structure of the data, the absolute coefficient sizes will be useful in determining what questions to use in the main study. It seems that the questions tapping into the research validity, quality and rigour are strongly correlated with each other; these questions will be used in the main study, along with the questions regarding the importance of the research.

Article	Cronbach's Alpha (Biology)	Cronbach's Alpha (Physics)
1	.820	.839
2	.875	.846
3	.877	.802
4	.778	.837
5	.897	.855
6	.853	.901
7	.846	.886
8	.861	.848
9	.847	.888
10	.893	.836

Table 26. Cronbach's Alpha values for each article, calculated using only the items tapping into the research validity, importance, quality, rigour and chance of dissemination for the pilot of Study 8.

	Validity	Importance	Quality	Rigour	Dissem.	Comp.
Validity	-	-.001	.930*	.865*	.452*	.555*
Importance	.429*	-	.076	.145	.440	-.158
Quality	.802*	.495*	-	.938*	.422	.439
Rigour	.709*	.448*	.788*	-	.249	.229
Dissemination	.473*	.497*	.499*	.399*	-	.660*
Comprehension	.431*	.241*	.433*	.341*	.425*	-

*Table 27. Pearson's correlation coefficients (*p<.05) between the evaluation items of interest for both the aggregated (top) and disaggregated data (bottom) for the pilot of Study 8.*

Main Study

1. Method

a. Participants

The sample size was based on obtaining 95% power to detect a small-to-medium effect size (d = 0.15) in a within-subjects t-test for comparing two face types. Participants were recruited online, using Amazon's MTurk, and 70 participants were excluded for recognizing either of the two articles they had read in the study.

The final sample, after exclusion, consisted of 558 participants (261 men, 297 women), with age ranging between 18 and 81 (M = 36.4, SD = 12.5). 94% of the participants reported being US nationals, and 96.5% had English as their first language. Participants were allocated to conditions in the following numbers: 150 in Male-Biology, 144 in Male-Physics, 129 in Female-Biology and 135 in Female-Physics.

b. Stimuli and Materials

The research summaries used in the main part of the study were selected from the articles used in the pilot study. Four biology articles and four physics articles were selected, based on their ratings on the composite measure comprised of the average of questions about validity, importance, quality and rigour from the pilot study. We chose articles with mean ratings around the middle of the scale; the articles were reasonably easy to comprehend, and did not have a recognition rate above 10% (Table 28). Two filler articles were also created, presenting the achievements and interests of athletes in first person, to simulate an "athlete profile" type of magazine section.

Biology Articles	Mean Quality	Mean Comprehension	Recognition
Study Suggests Earth Life Began on Mars	3.98	4.76	6.35%
Slime Mould Is Smarter Than You Think	4.39	4.97	4.76%
Beneath Pacific Lies Ancient, Barely Alive Bacteria	4.52	5.18	1.49%
Earth Holds 8.7M Species, and Most of Them are Still Undiscovered	4.52	5.17	4.76%
Physics Articles	Mean Quality	Mean Comprehension	Recognition
Dark Matter Particles Detected Deep in Mine	4.01	4.84	9.52%
Bloodhound Diary: It's rocket science	4.53	5.29	4.84%
World's Next Timekeeper: Quantum Superclock?	4.67	4.63	0%
Final chapter to be published, in decades-long Gravity Probe B project	4.69	3.91	1.56%

Table 28. Titles or articles used in Studies 8 and 9; Study 9 only used the Physics stories.

The photos were selected from the sample of photos used in Study 7. The highest rated two photos and lowest rated two photos from Study 7 were used, for both men and women, resulting in a total of 8 photos. Additionally, two male and two female athlete photos were collected as fillers, following a Google search for the terms “athlete”, “male athlete” and “female athlete”. The photos were cropped, edited and resized to match the main stimuli.

c. Design and Procedure

The study had a 2x2x2 mixed between- and within-subjects design, with face-type (high rated versus low rated) as a within-subjects factor, and gender (male photos vs. female photos) and discipline (biology articles vs. physics articles) as between-subjects factors. The articles were paired with the faces as follows: four versions of the task were created for each gender-by-discipline combination, using four 4x4 Latin-squares to ensure that the articles and faces were equally matched up. The participants were randomly allocated to one of the 16 conditions (4 gender-by-discipline combinations, each with its’ own 4x4 Latin square), and the program randomly selected one of the high-rated faces (with the article it was paired up with) and one of the low-rated faces (with the article it was paired up with) to be shown to the participant. The order in which these two face-article combinations were seen was randomised for each individual. The participants were then presented with the gender-matched filler articles (two athlete photos paired with sports-based articles). Finally, participants were prompted with the faces they had seen, as well as the titles of the article they had read (each face-title combination at a time), and were asked to rate the research they had read about on the following topics, on a 7-point Likert scale: validity (“How valid were this scientist's conclusions?”), quality (“How good was the overall quality of this scientist's research?”), rigour (“How rigorous was this scientist's research?”) and importance (“How important are this scientist's findings?”); we also checked whether they had previously read about the research outside of the survey. The order in which the two face-

title probes were presented was randomised, and the order of the evaluation questions was also randomised individually for each participant.

Demographic information was also collected from the participants, alongside information regarding their engagement with science (see Appendix A).

2. Results

a. Internal Reliability

Cronbach's Alpha values were computed, to establish whether the questions probing research validity, quality, rigour and importance could be collapsed into a single measure. This was done separately for the "low" face responses, and for the "high" face responses. The results suggested that the measures had high internal reliability, for both ratings of high (α = .887) and low (α = .874) faces. The four ratings were then averaged into a single measure of the overall goodness of the research, referred to as "Quality".

b. Mixed effects linear regression

The main question of interest was whether probing participants with a photo rated high on "looking like a good scientist" would elicit significantly higher ratings of the research that the participants read about, than probing them with a low-rated photo, and whether this differed depending on the discipline of the article read or the gender of the face presented, or any participant-level variables.

Using mixed effects regression (*lme4*, Bates et al., 2015), the overall rating of research quality ("Quality") was regressed onto the type of face the research was probed with (high or low; "Facetype"), the discipline of the article (biology or physics; "Discipline"), the gender of the scientist (male or female; "Gender"), all the possible 2-way and 3-way interactions between them, as well as participants' gender, age and science engagement. Considering the structure of the data, and that each participant rated two pieces of research (one

probed by a high face, the other probed by a low face), the model included a random intercept for participant and random, uncorrelated slopes for face-type, in line with previous analyses. The results indicated that research probed by high-rated faces (mean Quality rating = 4.93) was rated significantly higher in quality than research probed by low-rated faces (mean quality rating = 4.70), as illustrated in Figure 11.

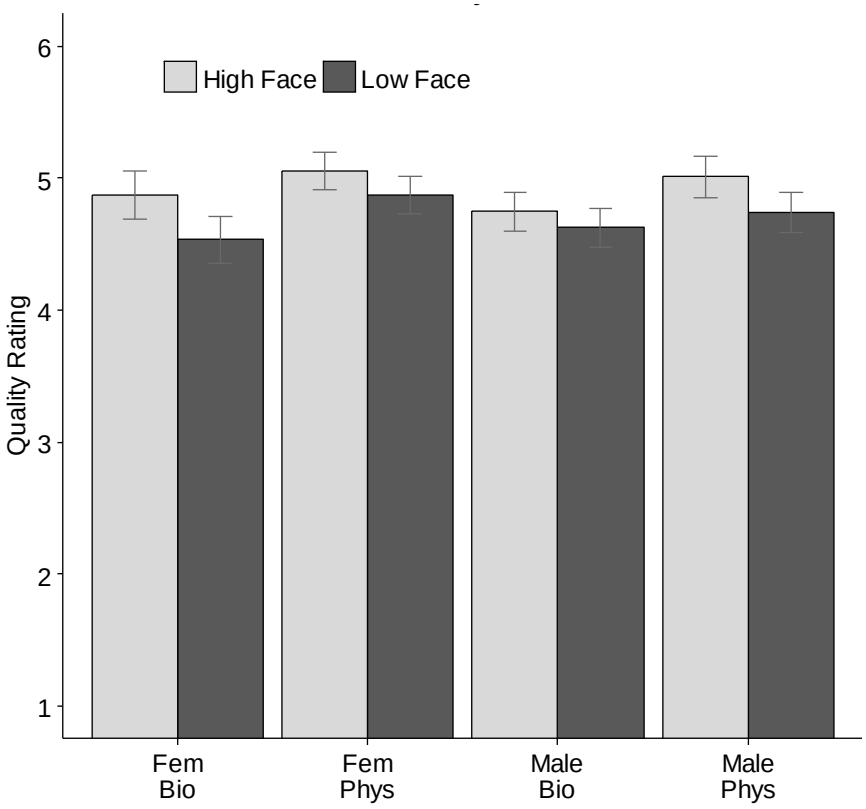


Figure 11. Mean ratings and error bars (Morey, 2008) on "Quality of Research" for research prompted by either high-rated faces or low-rated faces, in the four conditions of Study 8.

Additionally, physics research pieces were rated higher in quality than biology research, and participants with higher science engagement gave overall higher ratings on the quality of the research; no interactions between face-type, gender and discipline were significant (see Table 29 and Figure 12).

Predictor	B	95% CI Low	95% CI High	*p*
Intercept	4.805	4.714	4.896	<.001
Facetype	0.161	0.083	0.238	<.001
Discipline	0.119	0.028	0.210	.011
Gender	0.017	-0.074	0.108	.716
Facetype * Disc.	0.001	-0.076	0.078	.978
Facetype * Gender	0.024	-0.054	0.101	.550
Disc. * Gender	0.009	-0.082	0.101	.844
Facetype * Disc. * Gender	-0.051	-0.129	0.026	.193
Part gender	0.081	-0.014	0.175	.094
Part age	-0.068	-0.160	0.024	.150
Part science engagement	0.150	0.055	0.245	.002

Table 29. Coefficients and p-values for the fixed effects of the factors (face-type, discipline, gender and their interactions), and participant-level variables (gender, age and science engagement) for predicting the quality of the research in Study 8.

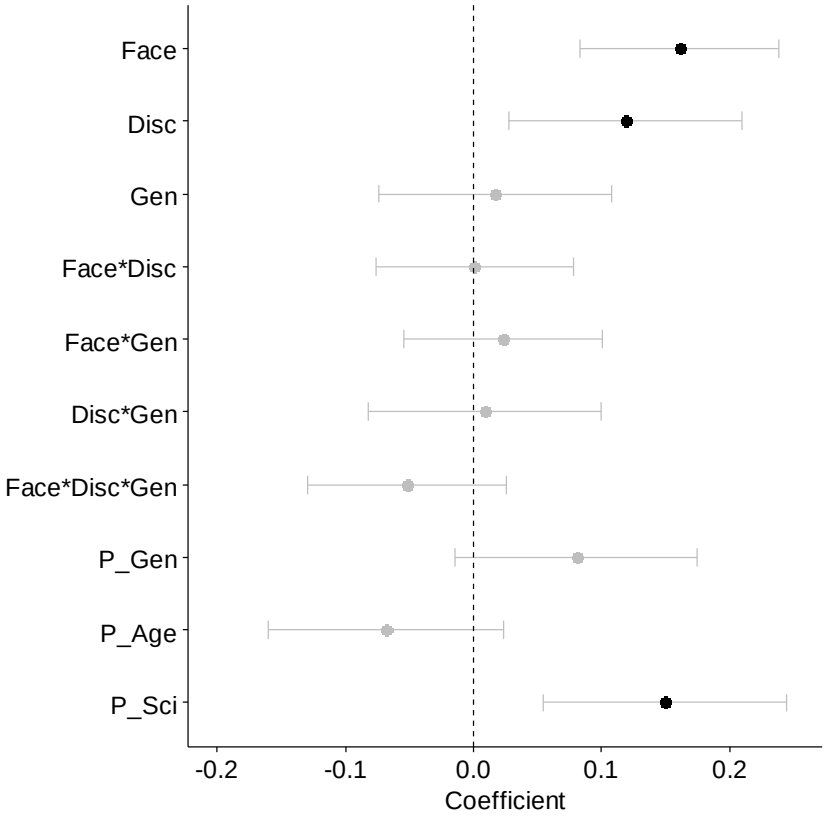


Figure 12. Regression coefficients and 95% Wald confidence intervals for the random intercept and random slopes model in Table 29. Coefficients with CIs that exclude zero are highlighted in black.

Discussion

Study 8 illustrates a potential bias in the way the public perceives scientific communications: research articles associated with photos of "good" scientists were rated significantly higher in quality than research articles associated and probed with "bad" scientists. Participants who felt more confident in their scientific knowledge and more engaged with science tended to judge the research articles to be higher in quality in general. These results suggest that the public is not only likely to extrapolate from the appearance of a scientist to their academic behaviour (Study 7), but they are likely to be influenced by the scientist's appearance when considering the quality of the research they are presenting.

Study 9: Is the public's opinion of scientific communications influenced by the perceived attractiveness and competence of the scientist?

In Study 9 we investigated whether a scientist's perceived attractiveness and competence predicts how people perceive the scientist's research. In Study 3, scientists who were perceived to be more facially competent and less attractive were perceived to look more like a good scientist. Given that research associated with "good" scientists was perceived as higher in quality, we expect research associated with highly competent but less attractive scientists will also be perceived as higher in quality. Study 9 was pre-registered on the Open Science Framework (osf.io/fterb).

Method

1. *Participants*

The minimum sample size of 800 participants was calculated based on 80% power to detect a small effect (0.1), with an alpha criterion of .05. The effect size was estimated using the effect size from Study 8 (d = 0.18). We are using a more conservative effect size here since the current manipulation is not as strong as the previous one: in Study 8, the photos differed at most by 4.20 standard deviation units on the "good" scientist measure, whereas the current photos differed at most by 3.66 standard deviation units on the same measure. Participants were recruited online using Amazon's MTurk; they were asked a simple memory/attention check after reading the science stories, and those who failed were re-directed to an "end of survey" page, being counted as non-completed. Three participants were excluded for reporting technical problems, and one was excluded for recognising all the articles. The discrepancy in the exclusion policy between this study (exclude participants who recognised all articles) and Study 8 (exclude participants who recognised any article) came from an error made when we submitted our preregistration for Study 9. We intended

for this study to have the same policy as Study 8, but decided it was best to keep to the publicly pre-registered plan for this study.

The final sample comprised of 824 participants (369 men, 455 women), with age ranging from 19 to 73 (M = 37.5, SD = 12.0). Approximately 89% of participants were US citizens, and 97.8% had English as their first language; participants were paid at the standard rate.

2. Stimuli and Materials

The research summaries used were the 4 physics summaries used in Study 8, matched in terms of quality ratings (Table 28). The scientist photos were the same as those used in Study 6 (8 photos of men, see Table 23), with two photos chosen for each cell of the design: high competence high attractiveness (HCHA), high competence low attractiveness (HCLA), low competence high attractiveness (LCHA), low competence low attractiveness (LCLA).

3. Design and Procedure

The study has a 2x2 within subjects design, with attractiveness (high or low) and competence (high or low) as within-subjects factors, as Study 6. Participants had one trial for each cell of the design.

The pairing of articles to cells of the design, the allocation of participants to versions and selection of photos was identical to the methodology of Study 6. After the presentation stage, an attention check was shown, asking participants to identify the topic which they had not read about, followed by the test phase (participants who failed this check were sent to an “end of survey” page). During the test phase, participants were prompted with the faces they had seen, as well as the titles of the articles they had read (each face-title combination at a time, in a randomized order), and asked to imagine they had been selected to judge how much each piece of research deserved to win a prize for excellence in science based on the following criteria: rigour of research, validity of conclusions, importance of

findings. Participants also provided an overall rating of how much the research deserved to win, and indicated whether they had read about the research before, or whether they had seen the scientist before.

Demographic information was also collected, alongside information regarding their engagement with science (see Appendix A).

Results

1. *Data Preparation*

Trials where participants claimed to recognise the research article or the researcher were excluded individually.

2. *Internal Reliability*

A Cronbach's Alpha analysis was conducted, to verify that the questions regarding the rigour, validity, importance, and overall deservingness of an award of the research could reasonably be collapsed into a single measure. Since each participant saw all 4 trials, nonrobust Cronbach's Alpha values were calculated separately for each type of trial (i.e., HCHA, HCLA, LCHA, LCLA). The results suggested that across all trials, the measures had high internal reliability (α_{HAHC} = .879, α_{HALC} = .872, α_{LAHC} = .871, α_{LALC} = .877), thus justifying averaging the four questions into a single measure assessing the overall quality of research.

3. *Mixed effects linear regression*

This project aimed to investigate whether a scientist's perceived competence and attractiveness would predict the quality ratings of the piece of research associated with the scientist. We were expecting research associated with scientists who are high in facial competence and low on physical attractiveness to be rated highest in quality, since these social dimensions are most strongly linked to looking like a "good scientist".

We conducted a mixed effects regression, using the *lme4* R package (Bates et al., 2015). Following our proposed analysis, the overall research quality rating was regressed onto the perceived competence of the scientist (high or low), their physical attractiveness (high or low), the interaction between them, the gender, age, and science engagement of the participant (i.e., participant-level variables), as well as the interactions between competence and participant-level variables, and attractiveness and participant-level variables; all interactions were computed manually. Given the nested structure of the data (four responses per participant) the model included a random intercept for participant, and random, uncorrelated slopes for competence, attractiveness and their interaction. Significance was evaluated using p-values produced using Satterthwaite approximations. The results suggest that research associated with more competent-looking scientists was rated higher in quality: the mean Quality rating for the LAHC group was the highest (4.99), followed closely by the mean Quality for HAHC (4.92), while the means for HALC (4.68) and LALC (4.68) were smaller (see Figure 13). The physical attractiveness of the scientist did not affect the quality ratings of the research, and the interaction between the two was not significant either (Figure 13).

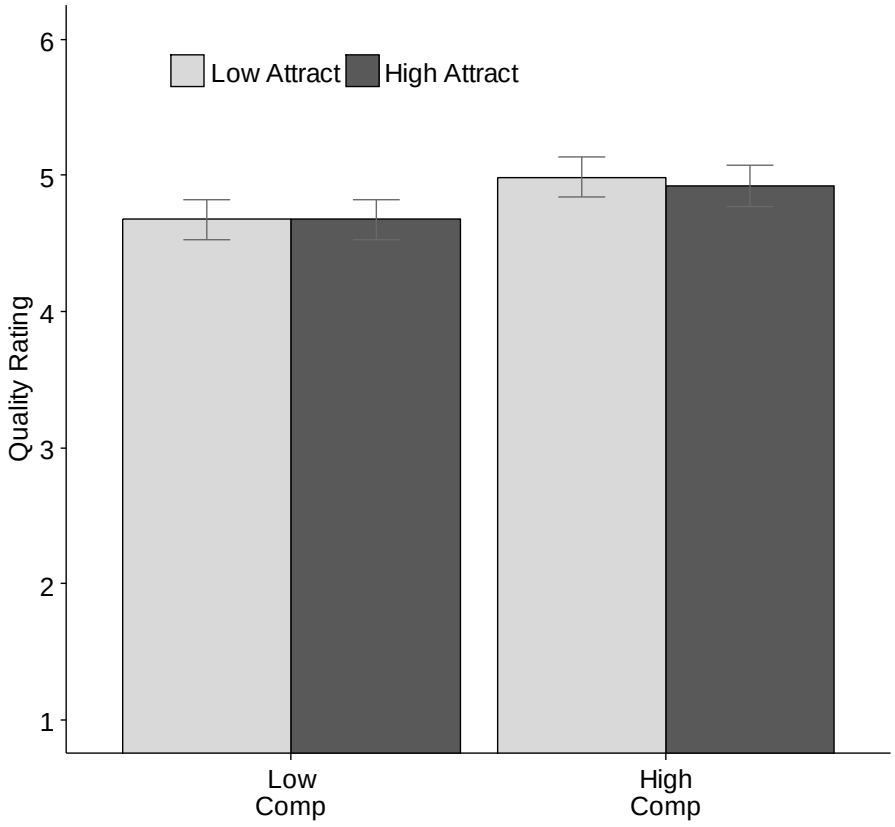


Figure 13. Mean quality of research ratings and error bars (Morey, 2008), for research associated with scientists that were either high or low on perceived competence and physical attractiveness in Study 9.

Participant-level variables significantly predicted the quality of research: women, younger participants and participants who were more highly engaged with science rated the research as higher in quality; no interactions between predictors and participant-level variables were significant (Table 30 and Figure 14).

Predictor	B	95% CI Low	95% CI High	*p*
Intercept	4.815	4.755	4.874	<.001
Competence	0.142	0.104	0.179	<.001
Attractiveness	-0.017	-0.053	0.020	.368
Comp. * Att.	-0.016	-0.052	0.021	.402
Part gender	0.102	0.041	0.163	.001
Part age	-0.080	-0.140	-0.020	.009
Part science engagement	0.094	0.033	0.155	.003
Comp. * Part gender	0.001	-0.037	0.040	.950
Comp. * Part age	0.013	-0.024	0.050	.497
Comp. * Part science engagement	0.037	-0.002	0.075	.060
Att. * Part gender	-0.010	-0.047	0.028	.610
Att. * Part age	0.006	-0.031	0.042	.758
Att. * Part science engagement	-0.022	-0.059	0.015	.252

Table 30. Coefficients, 95% Wald confidence intervals and p-values for the fixed effects of the factors (competence, attractiveness and their interaction), participant-level variables (gender, age and science engagement), as well as 2-way interactions between the factors and participant-level variables, for predicting the quality of the research in Study 9.

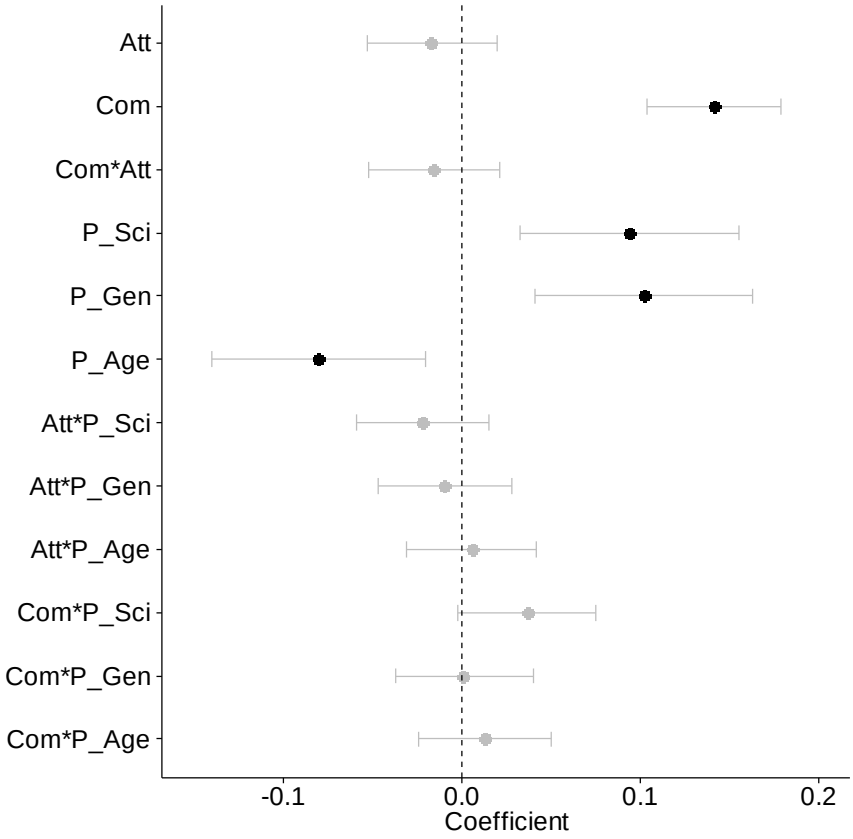


Figure 14. Regression coefficients and 95% Wald confidence intervals for the random intercept and random slopes model in Table 30. Coefficients with CIs that exclude zero are highlighted in black.

Discussion

Study 9 illustrated that scientific work associated with more competent looking scientists was judged to be more deserving of a prize, suggesting a potential bias in how scientific work is evaluated. The attractiveness of the scientist did not have an effect of the public's perception of research quality, and this is likely due to the weaker manipulation employed in this study. Similarly to Study 6, the ecological validity of the stimuli meant that the low-attractiveness photos differed from high-attractiveness photos by 0.86 points on the "good" scientist measure, as compared to low-competence and high-competence photos, which differed by 1.96 points on the "good" scientist measure. Using real photos of scientists led to a stronger competence manipulation, which was easier to detect than the attractiveness

manipulation. Regardless, Study 9 builds on the findings from Study 8, proposing important implications regarding the public's opinion of scientific news.

Chapter Summary

In sum, three studies examined how face-based impressions affected the evaluation of scientific news, and deservingness of positive scientific outcomes: scientists who looked more like a stereotypically "good" scientist were considered to be more deserving of positive scientific outcomes, while research paired with such scientists was judged to be higher in quality. Furthermore, more competent looking scientists were believed to produce more prize-worthy research. These findings support the idea that the public shares a stereotypical image of what a "good" scientist looks like, and believes that looking like a good scientist can have real-life implications. The fact that real science news stories were judged more favourably when arbitrarily associated with different faces suggests that facial cues are a potential source of bias in science communication.

CHAPTER 5: GENERAL DISCUSSION

Overview of findings; practical and theoretical implications

Across three studies, we found good agreement about what both a "good" scientist and a scientist likely to produce interesting research (i.e., a scientist whose work the public would be interested in learning about) looks like. A series of six further studies revealed that these first impressions based on the facial appearance of the scientist influenced both the public's choice between scientific news stories paired with photos of scientists, and their perception of the quality of the scientist's research. This project's findings are relevant not only to the social psychology of science communication and the development of the impression formation / person perception literature, but can also help inform the development of novel science communication policies.

Our studies indicated that first impressions of scientists were related to the basic social-psychological traits of competence, sociability, morality and attractiveness; we ensured that trait judgements were genuine first impressions and not based on knowledge about the person (c.f., Ballew & Todorov, 2007). Interestingly, some of the traits defining a "good" scientist were different from, if not even opposite to, the traits defining a scientist likely to produce interesting research. This trend suggests that initial interest in a scientist's work may require different qualities to inspiring a positive evaluation of the scientist's work. Members of both the UK and US public agreed on which social traits were more important in defining a "good" scientist and one likely to produce interesting research: people were more interested in learning about the research of scientists who looked more competent, moral and physically attractive, while scientists who looked more competent and moral, but less sociable and attractive were perceived as looking more like good scientists. Although competence and morality were desirable traits for both "good" scientists and scientists likely to produce interesting research, being physically attractive was beneficial for increased interest in one's scientific work, but detrimental for being perceived as conducting

good science. Competence was the dimension most consistently associated with both good scientists and interest in a scientist's research, providing support for competence being a core dimension of social judgement, with predictive power in many social outcomes (see Todorov et al., 2015, for a review). We have also found evidence that sociability and morality are different components of warmth, and that the three-factor model of social judgement (competence, sociability, morality) was a better fit to our data than the two-factor model (competence and warmth). Taken together, these results suggest that morality played a distinct role in impression formation in science communication, and should be considered a core dimension for most social relations (Brambilla & Leach, 2014; Goodwin, 2015). The detrimental effects of attractiveness are not fully unexpected, and have been identified for political outcomes (looking attractive was negatively correlated with winning elections; Mattes et al., 2010), and even legal outcomes (more attractive criminal defendants received harsher sentences, when the offence was attractiveness-related; Sigall & Ostrove, 1975). Moreover, the stereotypical view of a scientist is as someone embarked on a quest for truth, but who holds little personal appeal (Schinske et al., 2015). This stereotype is supported by the negative association between physical attractiveness and the perceived ability of conducting good research found in our studies.

In addition to what first impressions people form of scientists, we also investigated how these face-based impressions may affect the selection and evaluation of scientific news paired with photos of scientists. In terms of choice of scientific news, equally interesting communications paired with photos of scientists who looked more likely to produce interesting research were selected more often; participants also expressed more interest in scientific news stories associated with more competent-looking scientists. Similar results were found when looking at people's evaluations of scientific news stories: real science news stories were judged more favourably in terms of research quality when they were paired with photos of "good" scientists, which was also the case when the news stories

were paired with more competent-looking scientists. We investigated these effects for both male and female scientists, biology and physics news stories, text and video-based communications, and found consistent results across gender, discipline, and communication channel. Our findings support recent research showing that agentic traits (i.e., competence) are considered to be more important for success in science than communal traits (i.e., sociability and morality; Ramsey, 2017). We have found this to be the case not only for science, but for science communication as well. Furthermore, the influence of the scientist's appearance was not substantially modulated by the public's engagement with science, suggesting that trust in the source (in our case, the scientist) played a larger role than knowledge about the subject (cf. Connor & Siegrist, 2010).

One particularly interesting effect was observed in Studies 6 and 9, where competence-based stereotyping was pitted against attractiveness-based stereotyping in two critical experiments, and was shown to have a more robust effect. This poses an interesting problem in terms of this project's central hypothesis. We worked on the assumption that any social traits relevant to people's first impressions of scientists (i.e., competent-looking but unattractive scientists are perceived to look more like "good" scientists) may also have an influence on people's selection and evaluation of scientific communications. However, despite finding a strong effect of attractiveness in defining people's perception of good scientists and scientists likely to produce interesting research, there was no such effect of attractiveness on people's selection (Study 6) and evaluation (Study 9) of science. While this could potentially be due to a weaker manipulation (as discussed previously), it could also suggest a bigger problem: despite both competence and attractiveness playing a strong role in defining people's perceptions of scientists, competence may have a stronger effect when it comes to making a decision about which scientific communication to engage with, or expressing your opinion about the communication, rending the effect of attractiveness null. It is also worth mentioning that other traits may have been confounded with competence

and attractiveness: low competence, high attractiveness faces were particularly young (mean age 26.07, compared to mean ages for the other groups which ranged from 42.02 to 52.62), while high competence faces were also particularly high on "looking like a good scientist" (mean good scientist ratings 7.16 and 6.06, compared to mean good scientist ratings of 4.96 and 4.34 for the low competence faces). Since these other traits were not included in our analysis, which was centred around attractiveness and competence, it is possible that they may have driven part of the effect; we suggest that this matter should be investigated further, for clarification.

Our findings speak also to the claims of the persuasion models discussed previously (i.e., ELM, Petty and Caccioppo, 1986 and HSM, Chaiken & Trope, 1999), as well as to stereotyping models, such as Fiske and Neuberg's (1990) continuum of impression formation model. According to the persuasion models, more in-depth processing of the information available is more likely to occur when the decision is personally relevant to the person, as this would motivate the individual to engage in systematic processing (Todorov, Chaiken & Henderson, 2002). Similarly, the continuum of impression formation model posits that, during the impression formation process, people make an initial categorization based on appearance and superficial features (e.g., gender, skin-colour). If they deem the target to be personally relevant, then they will try to move past category based impressions and focus more on individuating features (Fiske & Neuberg, 1990). These theoretical models hold implications for the differences in the effects found across our studies. For example, in the three face-rating studies, there is evidence of category-based processing: the top photos rated highly on "looking like a good scientist" resembled the stereotypical image of a scientist (older, male, with white hair and glasses). This suggests that, without any other information available and with no personal relevance, people were relying on a salient exemplar of the scientist category to make their judgements. A similar effect is present in Study 7: when participants were asked to make a decision on which scientist was more likely

to find themselves in a positive or negative research scenario, the only information available was the scientist's appearance; furthermore, the decision had no personal implication for the participants, thus resulting in a large effect of "looking like a good scientist". Conversely, in Study 8, faces of scientists were paired with scientific articles; there was still an effect of "looking like a good scientist", but it was considerably smaller. This difference can be explained from the perspective of persuasion and stereotyping models: the design of Study 8 offers more information available to the participants to use when making decisions, thus resulting in a smaller effect of superficial features. Moreover, engaging with a scientific article makes the decision more personally relevant, and thus making participants more likely to process the information systematically. This is reflected in the participant's science engagement score predicting their quality of rating scores, and also in a stronger effect of "looking like a good scientist" for physics articles (which, one could argue, discuss topics less relevant to a general audience). The effect differences between studies involving photos only, and studies involving photos and additional information suggest that persuasion and stereotyping models have implications for impression formation in science communication, as discussed above.

Across our studies we provided evidence that the same piece of research will be evaluated differently depending on the appearance of the scientist the research was arbitrarily paired with. This is particularly noteworthy considering the recent arguments against the phenomenon called "face-ism": Olivola et al. (2014) claim that the social inferences people make based only on facial appearance are inaccurate and unreliable, and thus we should stop people from using inferences made from faces as the basis for important social decisions. This claim is supported by further evidence suggesting that different facial images of the same person can lead to different impressions, and that the preference for images depends on the context (Todorov & Porter, 2014). Despite the potential negative consequences of face-ism, there is a kernel of truth in facial judgements, and even small

effects (just above the level of random guessing) can be of importance, suggesting that people have the ability to detect social information from faces (Bonnefon, Hopfensitz & De Neys, 2014). The consistency of our results indicates that facial cues are a potential source of bias in science communication; although this bias was not always large, its practical significance cannot be understated given that scientific findings increasingly shared and disseminated via web-based platforms (e.g., TED Talks). For example, the 60% preference for finding out more about research associated with scientists likely to produce interesting research we identified in the Video condition of Study 5 would translate into hundreds of thousands of extra views and shares on platforms like YouTube, Facebook, Twitter, and TED Talks. Considering the combination of a particularly strong effect for video communications, and the increasing use of videos and media to inform the public of scientific findings, judgements based on the facial appearance of the scientist are increasingly likely to influence the public's interaction and engagement with available scientific research.

Throughout this project, we focused on biology and physics, as two different sides of science: both disciplines being stereotypically considered "science" (i.e., wearing lab coats and conducting experiments), with biology being slightly on the "softer" side of science. We argue that our findings are generalizable to a larger category of scientific disciplines, but chose not to include others due to the worry that human-focused sciences such as psychology and sociology might not be part of the category exemplars of "science" for the majority of the public. This would not be surprising, particularly given the stereotypical example of what a "good" scientist looks like (older male with white hair and wearing a lab coat, as discussed in Study 1). Attractive-based stereotyping proposes a "halo" effect, where physically attractive people are also perceived to possess other socially desirable traits (Eagly et al., 1991). However, this was not necessarily the case for scientists in biology and physics, as illustrated in our studies: attractiveness had a positive role in defining scientists who looked likely to produce interesting research, but a negative role for in defining "good"

scientists. Furthermore, the importance of attractiveness diminished when participants were asked to make a choice or voice their opinion about scientific research communications, indicating that attractiveness may play a more dominant role when people are asked to consider what traits define a scientist. These findings fall in line with previous research having shown that physical attractiveness is valued in communicators (Mendez & Mendez, 2016): scientists who looked more attractive were perceived to receive more interest in their research, suggesting that attractiveness is something the public looks for in a scientific communicator. Dilger et al. (2015) found that attractiveness did not predict research success, and our findings speak to that: more attractive scientists were less likely to be seen as "good" scientists. Our results also mirror those of Talamas et al. (2016) who found that attractiveness did not predict actual academic performance, but it did predict perceived academic performance; similarly, attractiveness predicted perceptions of scientists in our project. It would be worthwhile to investigate whether attractiveness predicts actual research success in the future.

Effects of gender differences in target and participant gender

Throughout this project, we tested the potential role of gender differences in science communication. Because our samples of scientist photos were collected from real University websites, the larger number of men compared to women in the both the UK and US samples towards males reflects a known gender imbalance in academia. Looking at the number of women at Undergraduate level in STEM (Botcherby & Buckner, 2012) comparatively to the number of women in our samples, one could conclude that there is a failure in either the recruitment or retention on permanent academic scientists, leading to the observed gender imbalance. Despite the male-dominated samples, our analyses revealed little gender differences in "good" scientist ratings – i.e., male and female scientists were rated similarly on whether they looked like scientists conducting good research. It is important to note that

the lack of gender differences was found in analyses taking into account all the other traits and dimensions in our study.

Across the nine studies, there were both null and significant effects of the target's (i.e., the scientist's) gender. In studies 1 and 3, the target's gender was positively correlated with attractiveness, suggesting that women in scientist samples were considered more attractive. Although there were no significant effects in the mixed effect regression, in both studies the target's gender was negatively correlated with "looking like a good scientist", but positively correlated with "looking likely to produce interesting research". In studies 4 and 5, scientist's gender was negatively related to the percentage of "interesting" scientist choices (i.e., women chosen less often), while in studies 7 (negative) and 8 (positive) there was no consistent effect of the target's gender on the quality of the article. These trends suggest that not only were female scientists perceived to be more attractive, but on a basic level, women were also perceived to look less like good scientists, an effect potentially related to their attractiveness. However, it is important to note that, in our analyses, gender effects were considered simultaneously with other effects, thus controlling for the other traits included in the analysis. Therefore, there may be net effects of gender if gender was to be considered in isolation, resulting in women being potentially judged less favourably than men (as seen in the zero-order correlations in Studies 1 and 3). The perception of female scientists as more attractive (and, in turn, looking less like good scientists) may be related to the pressure women experience to present themselves in a favourable light in their professional careers: the photos of women were consistently of a higher standard than men's, with women appearing to have put more effort into their appearance. One way to address these naturally-occurring differences would be to adopt some University-level standardization of staff photos, minimizing the differences between the effort men and women put into their appearance.

Although we did not directly investigate the stereotypes of scientists as men, our findings provide some evidence for the stereotypical view of scientists as men (Nosek et al., 2009), as highlighted by the zero-order correlations (women seen as more attractive, more interesting, but looking less like a good scientist, which agrees with the literature on sexism in academic careers). Furthermore, our analyses revealed little evidence for any effect of gender on either the selection or evaluation of scientific communications, going against research suggesting that publication success is biased towards women (Budden et al., 2008).

Moreover, the effects of gender (or lack thereof) may depend on the population the faces were sampled from: in the only study that used photos from a standardised face database (as opposed to photos of real scientists), women were rated lower on looking like a "good" scientist. The sample was also comprised of an equal proportion of each gender, suggesting that gender effects in science communication should be investigated further.

Our investigation was not limited to the scientist's gender: we also collected information about the participant's gender, and its predictive value. For the face-rating studies, there was no consistent pattern: participant gender was negatively related to both ratings of looking like a good and an interesting scientist in Study 1, negatively related to ratings of good scientist, but positively related to ratings of interesting scientist in Study 2, and positively related to both in Study 3. In studies 4 and 5, participant gender negatively predicted choosing interesting scientists, while in study 6 it has an overall positive effect, a negative interaction with competence but a positive interaction with attractiveness. In study 7, participant gender was negatively related to choosing "good" scientists in "good" scenarios; in study 8, it was positively related to ratings of research quality, while in study 9 it positively predicted ratings of research quality, having a null interaction with competence, but a negative interaction with attractiveness. Across the studies, there is not straightforward trend suggesting that either male or female participants were harsher or

more lenient, which goes against the literature on sexism in scientific careers. Women tend to be judged more harshly (Caleo, 2016), while women have been found to think they are less able than men, even when this was not the case (Spencer, Steele & Quinn, 1999). Although we did not find evidence for such effects here, they are still a prominent issue for women in STEM, potentially acting like a self-fulfilling prophecy (girls considering themselves less able at STEM subjects, and in turn pursuing them less). As a result, we recommend pursuing a further investigation of both target and participant gender effects in science communication.

Limitations and future directions

We focussed on scientists and participants from the US and UK, and found consistent results across the two nationalities. Although there is evidence suggesting a certain universality of social dimensions perception across Western and Eastern cultures (e.g., Walker et al., 2011), it is possible that social roles, expectations and stereotypes of scientists will differ depending on cultural and societal standards. For example, Rule et al. (2010) found high agreement in face ratings of electoral candidates between American and Japanese participants; however, the actual traits that predicted electoral success differed depending on culture. Thus, it would not be surprising if future research found some agreement between cultures on the social dimensions extracted from faces of scientists, with possible discrepancies about the social traits relevant in predicting scientific popularity and achievement.

In each study, we measured participants' self-reported engagement with science, tapping into both their interest in and their knowledge of science. Although this captures some information regarding people's feeling about science, it would be important to probe further the public's knowledge and attitudes towards scientific research in the context of face-based impressions of their work. One issue we did not address was the beliefs people

had about the research topics we used in our studies; Kahan (2010) argued that people will be biased in their interpretation of new information, in a way that reinforces their beliefs and predispositions. Different groups with different cultural backgrounds (e.g., individualistic vs. collectivistic cultures) will perceive evidence more favourably when it reinforces their outlooks, or when it comes from experts who share their values (Kahan, 2010), and this should be considered when working on improving science communication. Moreover, even individual differences between the participants' trait empathy and numeracy level can affect how scientific information is perceived, and which type of information is processed better (Knobloch-Westerwick, Johnson, Silver & Westerwick, 2015), so future research should take into account the public's knowledge and attitudes towards science.

One of our goals for this project was to use ecologically valid stimuli as the basis of social judgements, so we collected photos of actual scientists, in an attempt to replicate the impression formation processes occurring in real-life. However, it would be informative to verify whether our findings replicate with artificially-constructed stimuli, such as computer-generated faces whose traits can be systematically manipulated (Todorov et al., 2015). This approach would be particularly useful for studies using extreme stimuli (i.e., high on competence, but low on attractiveness). For example, our face-rating data suggested an opposite effect of attractiveness on measures of good scientist and interesting research: more attractive scientists were rated more likely to receive interest in their work, while less attractive scientists were considered to look more like scientists conducting good research. From these findings, one could reasonably expect attractiveness to lead to increased research interest, but decreased perceived quality. However, when we manipulated attractiveness and competence by choosing extreme stimuli, we found little evidence for the expected effect of attractiveness – we hypothesized that the failure to find an effect was due to rather small differences in ratings of good scientists and interesting research

between the low- and high-attractiveness stimuli used. Using computer-generated stimuli should address this problem, by allowing us to build stimuli that are tailored to our research needs – we would be able to construct photos that have a better separation on the social dimensions of interest. This could be achieved using either computer-generated faces (Todorov, Said, Engell & Oosterhof, 2008), or custom-modified real photos (Walker & Vetter, 2009), allowing a better test of whether the same trait can increase initial engagement with a scientist's work, and decrease the perceived quality of the research associated with the same scientist.

Finally, we considered the two- and three-factor models of social judgement, with a particular emphasis on the three-factor model of competence, sociability and morality. This choice was based on the expectation that the three social dimensions would be key to investigating the facets of science communication we addressed in this project. Despite the evidence we found for the contributions of first impressions of competence, sociability and morality on science communication, there are other models of social judgement which we did not address. For example, Sutherland et al. (2015) investigated personality judgements of the Big Five model extracted from every day, ecologically valid photos of faces, and were able to create models of facial attributes that successfully predicted extraversion, agreeableness and openness to experience. Moreover, Koch, Imhoff, Dotsch, Unkelbach and Alves (2016) applied a different three-factor model to group stereotypes: the researchers looked at agency/socioeconomic success, conservative-progressive beliefs and communion (ABC), claiming to have addressed a gap in the popular warmth/communion – competence/agency model. Koch et al. (2016) argued that by constraining participant ratings to a set of pre-agreed dimensions, one would not tap into the dimensions spontaneously considered when judging social groups. While Koch et al.'s (2016) model was not really appropriate for this project's data (the trait ratings we employed do not map onto the ABC model), and it might be better suited for describing broad social groups rather than

individuals, it does raise an interesting research question. Thus, future research should consider other models of social judgement, and how well first impressions on these traits map onto real consequences of science communication.

Conclusion

To conclude, we investigated what first impressions people form based on a scientist's appearance, and whether these impressions are likely to influence the selection and evaluation of scientific communications to the general public. Our findings portray science as a social activity, whose outcomes are likely to depend on the facial appearance of the scientist. At a societal level, the influence of facial appearance on scientific communications is potentially dangerous, as it risks biasing both public attitudes and government actions involved with scientific issues of importance to society, such as climate change (Somerville & Hassol, 2011). Furthermore, there are implications for the scientists themselves: good, efficient communication between the scientists and the public has been shown to increase academic performance (Jensen, Rouquier, Kreimer & Croissant, 2008). Thus, first impressions based on the facial appearance of the scientist are not only likely to bias the popularity and level of acceptance of a scientist's work among the general public, but may go as far as influencing which scientific research is funded, conducted, and by whom.

REFERENCES

2011 Census, Key Statistics and Quick Statistics for local authorities in the United Kingdom - Part 1. (2013, October 11). *Office for National* Statistics. Retrieved August 27, 2015, from http://www.ons.gov.uk/ons/rel/census/2011-census/key-statistics-and-quick-statistics-for-local-authorities-in-the-united-kingdom---part-1/index.html.

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of personality and social psychology*, *93*(5), 751-763.

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in experimental social psychology*, *32*, 201-271.

Ambady, N., & Gray, H. M. (2002). On being sad and mistaken: mood effects on the accuracy of thin-slice judgments. *Journal of personality and social psychology*, *83*(4), 947-961.

Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness. *Journal of Consumer Psychology*, *16*(1), 4-13.

Ambady, N., LaPlante, D., & Johnson, E. (2001). Thin-slice judgments as a measure of interpersonal sensitivity. In *Interpersonal sensitivity: Theory and measurement.* (pp. 89-101). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness.*Journal of personality and social psychology*, *64*(3), 431-441.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, *104*(46), 17948-17953.

Banchefsky, S., Westfall, J., Park, B., & Judd, C. M. (2016). But You Don't Look Like A Scientist!: Women Scientists with Feminine Appearance are Deemed Less Likely to be Scientists. *Sex Roles*, 1-15.

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.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological science*, *15*(10), 674-679.

Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2015). Face-ism and kernels of truth in facial inferences. *Trends in cognitive sciences*, *19*(8), 421-422.

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, *62*(4), 645-657.

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of personality and social psychology*, *86*(4), 599-614.

Botcherby, S., & Buckner, L. (2012). Women in Science, Technology, Engineering and Mathematics: from Classroom to Boardroom. *UK Statistics, London*.

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. *Social Cognition*, *32*(4), 397-408.

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology, 41*(2), 135-143.

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V. Y. (2012). You want to give a good impression? Be honest! Moral traits dominate group impression formation. *British Journal of Social Psychology*, *51*(1), 149-166.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Budden, A. E., Tregenza, T., Aarssen, L. W., Koricheva, J., Leimu, R., & Lortie, C. J. (2008). Double-blind review favours increased representation of female authors. *Trends in ecology & evolution*, *23*(1), 4-6.

Caleo, S. (2016). Are organizational justice rules gendered? Reactions to men's and women's justice violations. *Journal of Applied Psychology, 101(*10), 1422-1435.

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, *41*(5), 1054-1072.

Castelli, L., Carraro, L., Ghitti, C., & Pastore, M. (2009). The effects of perceived competence and sociability on electoral outcomes. *Journal of Experimental Social Psychology*, *45*(5), 1152-1155.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. Guilford Press. Chicago

Chiao, J. Y., Bowman, N. E., & Gill, H. (2008). The political gender gap: Gender bias in facial inferences that predict voting behavior. *PLoS One*, *3*(10), e3666.

Connor, M., & Siegrist, M. (2010). Factors influencing people's acceptance of gene technology: The role of knowledge, health expectations, naturalness, and social trust. *Science communication*, *32*(4), 514-538.

Corbett, J. B., & Durfee, J. L. (2004). Testing public (un) certainty of science media representations of global warming. *Science Communication*, *26*(2), 129-151.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology*, *40*, 61-149.

Das, S. K. (2013). Scientific communication: Understanding scientific journals and articles. *Global Media Journal - Indian Edition, 4*(1), 1-10.

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and social psychology bulletin*, *26*(10), 1171-1188.

Diekman, A. B., & Goodfriend, W. (2006). Rolling with the changes: A role congruity perspective on gender norms. *Psychology of Women Quarterly*,*30*(4), 369-383.

Dilger, A., Lütkenhöner, L., & Müller, H. (2015). Scholars' physical appearance, research performance, and feelings of happiness. *Scientometrics*, *104*(2), 555-573.

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562-571.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological bulletin*, *110*(1), 109-128.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychological Science*, *17*(5), 383-386.

Farenga, S. J., & Joyce, B. A. (1999). Intentions of young students to enroll in science courses in the future: An examination of gender differences. *Science Education*, *83*(1), 55-75.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77-83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, *82*(6), 878-902.

Fiske, S. T., & Dupree, C. (2014). Gaining trust as well as respect in communicating to motivated audiences about science topics. *Proceedings of the National Academy of Sciences of the United States of America*, 13593-13597.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in experimental social psychology*, *23*, 1-74.

Fiske, S. T., Xu, J., Cuddy, A. C., & Glick, P. (1999). (Dis) respecting versus (dis) liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of Social Issues*, *55*(3), 473-489.

Flaherty, D. K. (2011). The vaccine-autism connection: a public health crisis caused by unethical medical practices and fraudulent science. *Annals of Pharmacotherapy*, *45*(10), 1302-1304.

Francis, G. (2015). Excess success for three related papers on racial bias. *Frontiers in psychology, 6*, 512.

Fuegen, K., Biernat, M., Haines, E., & Deaux, K. (2004). Mothers and fathers in the workplace: how gender and parental status influence judgments of job-related competence. *Journal of Social Issues*, *60*(4), 737-754.

Gabriel, U., Gygax, P., Sarrasin, O., Garnham, A., & Oakhill, J. (2008). Au pairs are rarely male: Norms on the gender perception of role names across English, French, and German. *Behavior Research Methods*, *40*(1), 206-212.

Gheorghiu, A. I., Callan, M. J., & Skylark, W. J. (2017). Facial appearance affects science communication. *Proceedings of the National Academy of Sciences, 114,* 5970-5975. doi: 10.1073/pnas.1620542114

Glanz, K., & Bishop, D. B. (2010). The role of behavioral science theory in development and implementation of public health interventions. *Annual review of public health*, *31*, 399-418.

Godfrey-Smith, P. (2003). *Theory and Reality: An Introduction to the Philosophy of Science*. University of Chicago Press.

Goodwin, G. P. (2015). Moral Character in Person Perception. *Current Directions in Psychological Science*, *24*(1), 38-44.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of personality and social psychology*, *106*(1), 148-168.

Gorn, G. J., Jiang, Y., & Johar, G. V. (2008). Babyfaces, trait inferences, and company evaluations in a public relations crisis. *Journal of Consumer Research*, *35*(1), 36-49.

Hall, J. A., Andrzejewski, S. A., Murphy, N. A., Mast, M. S., & Feinstein, B. A. (2008). Accuracy of judging others' traits and states: Comparing mean levels across tests. *Journal of Research in Personality*, *42*(6), 1476-1489.

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.

Hartley, J. (2003). Improving the Clarity of Journal Abstracts in Psychology The Case for Structure. *Science Communication*, *24*(3), 366-379.

Hartz, J., & Chappell, R. (1997). *Worlds apart: How the distance between science and journalism threatens America's future*. Nashville, TN: First Amendment Center.

Haynes, R. (2003). From alchemy to artificial intelligence: Stereotypes of the scientist in Western literature. *Public Understanding of Science*, *12*(3), 243-253.

Heflick, N. A., Goldenberg, J. L., Cooper, D. P., & Puvia, E. (2011). From women to objects: Appearance focus, target gender, and perceptions of warmth, morality and competence. *Journal of Experimental Social Psychology*, *47*(3), 572-581.

Herbert, J., & Stipek, D. (2005). The emergence of gender differences in children's perceptions of their academic competence. *Journal of Applied Developmental Psychology*, *26*(3), 276-295.

Illes, J., Moser, M. A., McCormick, J. B., Racine, E., Blakeslee, S., Caplan, A., ... & Weiss, S. (2009). Neurotalk: improving the communication of neuroscience research. *Nature Reviews Neuroscience*, *11*(1), 61-69.

Imhoff, R., & Koch, A. (2017). How Orthogonal Are the Big Two of Social Perception? On the Curvilinear Relation Between Agency and Communion. *Perspectives on Psychological Science*, *12*(1), 122-137.

Jensen, P., Rouquier, J. B., Kreimer, P., & Croissant, Y. (2008). Scientists who engage with society perform better academically. *Science and public policy*, *35*(7), 527-541.

Joo, J., Steen, F. F., & Zhu, S. C. (2015). Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3712-3720).

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *Journal of personality and social psychology*, *89*(6), 899-913.

Kahan, D. (2010). Fixing the communications failure. *Nature*, *463*(7279), 296-297.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kareklas, I., Muehling, D. D., & Weber, T. J. (2015). Reexamining Health Messages in the Digital Age: A Fresh Look at Source Credibility Effects.*Journal of Advertising,* 2015 Forthcoming.

Kervyn, N., Bergsieker, H. B., & Fiske, S. T. (2012). The innuendo effect: Hearing the positive but inferring the negative. *Journal of Experimental Social Psychology*, *48*(1), 77-85.

Kirkman, J., & Turk, C. (2002). *Effective writing: improving scientific, technical and business communication*. Taylor & Francis.

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. *Science Communication*, *35*(5), 603-625.

Knobloch-Westerwick, S., Johnson, B. K., Silver, N. A., & Westerwick, A. (2015). Science exemplars in the eye of the beholder: How exposure to online science information affects attitudes. *Science Communication*, *37*(5), 575-601.

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of personality and social psychology*, *110*(5), 675-709.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmerTest'. *R package version*, 2-0.

Landy, J. F., Piazza, J., & Goodwin, G. P. (2016). When It's Bad to Be Friendly and Smart The Desirability of Sociability and Competence Depends on Morality. *Personality and Social Psychology Bulletin*, *42*(9), 1272-1290.

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: the importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of personality and social psychology*, *93*(2), 234-249.

Lenz, G. S., & Lawson, C. (2011). Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science*, *55*(3), 574-589.

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior*, *28*(1), 18-27.

Logan, R. A. (2001). Science Mass Communication Its Conceptual History. *Science Communication*, *23*(2), 135-163.

Martinez-Conde, S. (2016). Has Contemporary Academia Outgrown the Carl Sagan Effect?. *Journal of Neuroscience*, *36*(7), 2077-2082.

Mattes, K., Spezio, M., Kim, H., Todorov, A., Adolphs, R., & Alvarez, R. M. (2010). Predicting election outcomes from positive and negative trait assessments of candidate images. *Political Psychology*, *31*(1), 41-58.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, *52*(1), 81-90.

McLean, H. M., & Kalin, R. (1994). Congruence between self-image and occupational stereotypes in students entering gender-dominated occupations. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *26*(1), 142-162.

Mead, M., & Metraux, R. (1957). Image of the Scientist among High-School Students. *Science*, *126*, 384-390.

Mendez, J. M., & Mendez, J. P. (2016). Student inferences based on facial appearance. *Higher Education*, *71*(1), 1-19.

Miller, A. G. (1970). Role of physical attractiveness in impression formation. *Psychonomic Science*, *19*(4), 241-243.

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 630-633.

Moore, A. (2006). Bad science in the headlines. *EMBO reports*, *7*(12), 1193-1196.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*(2), 61-64.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474-16479.

Murphy, N. A., Hall, J. A., & Colvin, C. R. (2003). Accurate intelligence assessments in social interactions: Mediators and gender effects. *Journal of personality*, *71*(3), 465-493.

Murphy, N. A., Hall, J. A., Schmid, M. M., Ruben, M. A., Frauendorfer, D., Blanch-Hartigan, D., ... & Nguyen, L. (2015). Reliability and validity of nonverbal thin slices in social interactions. *Personality & social psychology bulletin*, *41*(2), 199-213.

National University Rankings. (2014). *National University Rankings.* In U.S. News & World Report. Retrieved September 5th, 2014 from http://colleges.usnews.rankingsandreviews.com/best-colleges/rankings/national-universities/data/spp%2B50

Naylor, R. W. (2007). Nonverbal cues-based first impressions: Impression formation through exposure to static images. *Marketing Letters*, *18*(3), 165-179.

Nelkin, D. (1995). *Selling science: How the press cover science and technology*. Rev. Ed. New York: Freeman.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... & Kesebir, S. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593-10597.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566-570.

Olivola, C. Y., & Todorov, A. (2010a). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*(2), 315-324.

Olivola, C. Y., & Todorov, A. (2010b). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*(2), 83-110.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087-11092.

Pagliaro, S., Brambilla, M., Sacchi, S., D'Angelo, M., & Ellemers, N. (2013). Initial impressions determine behaviours: Morality predicts the willingness to help newcomers. *Journal of Business Ethics*, *117*(1), 37-44.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, *162*(1), 8-13.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in experimental social psychology*, *19*, 123-205.

Poutvaara, P., Jordahl, H., & Berggren, N. (2009). Faces of politicians: Babyfacedness predicts inferred competence but not electoral success. *Journal of Experimental Social Psychology*, *45*(5), 1132-1135.

Projet Shtooka. (n.d.). *Welcome (English)*. Retrieved April 22, 2015, from http://shtooka.net/listen/eng/welcome

Qualtrics, L. L. C. (2014). Qualtrics [software].

Ramsey, L. R. (2017). Agentic traits are associated with success in science more than communal traits. *Personality and Individual Differences*, *106*, 6-9.

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PloS one*, *7*(3), e34293.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annu. Rev. Psychol.*, *57*, 199-226.

Rosseel, Y., Oberski, D., & Byrnes, J. (2011). Lavaan: Latent variable analysis. *R package version 0.4-11*.

Rule, N. O., & Ambady, N. (2008). The face of success inferences from chief executive officers' appearance predict company profits. *Psychological Science*,*19*(2), 109-111.

Rule, N. O., & Ambady, N. (2010). First impressions of the face: Predicting success. *Social and Personality Psychology Compass*, *4*(8), 506-516.

Rule, N. O., Ambady, N., Adams Jr, R. B., Ozono, H., Nakashima, S., Yoshikawa, S., & Watabe, M. (2010). Polling the face: prediction and consensus across cultures. *Journal of personality and social psychology*,*98*(1), 1-15.

Satterfield, J. M., Spring, B., Brownson, R. C., Mullen, E. J., Newhouse, R. P., Walker, B. B., & Whitlock, E. P. (2009). Toward a Transdisciplinary Model of Evidence-Based Practice. *The Milbank Quarterly, 87*(2), 368–390. http://doi.org/10.1111/j.1468-0009.2009.00561.x

Savić, J. (2003). Effective scientific communication in biomedicine. *Archive of Oncology*, *11*(3), 201-202.

Schinske, J., Cardenas, M., & Kaliangara, J. (2015). Uncovering scientist stereotypes and their relationships with student race and student success in a diverse, community college setting. *CBE-Life Sciences Education*, *14*(3), ar35.

Seymour, B., & Vlaev, I. (2012). Can, and should, behavioural neuroscience influence public policy?. *Trends in cognitive sciences*, *16*(9), 449-451.

Shapin, S. (1996). *The scientific revolution*. University of Chicago Press.

Shonkoff, J. P., & Bales, S. N. (2011). Science does not speak for itself: Translating child development research for the public and its policymakers. *Child Development*, *82*(1), 17-32.

Siegrist, M., & Cvetkovich, G. (2000). Perception of hazards: The role of social trust and knowledge. *Risk analysis*, *20*(5), 713-720.

Sigall, H., & Ostrove, N. (1975). Beautiful but dangerous: Effects of offender attractiveness and nature of the crime on juridic judgment. *Journal of Personality and Social Psychology*, *31*(3), 410-414.

Skylark, W. J. & Gheorghiu, A. I. (2014). Unpublished data.

Sofer, C., Dotsch, R., Wigboldus, D. H., & Todorov, A. (2015). What Is Typical Is Good The Influence of Face Typicality on Perceived Trustworthiness. *Psychological Science*, *26*(1), 39-47.

Somerville, R. C., & Hassol, S. J. (2011). Communicating the science of climate change. *Physics Today*, *64*(10), 48-53.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology, 35*(1), 4-28.

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105-118.

Sutherland, C. A., Rowley, L. E., Amoaku, U. T., Daguzan, E., Kidd-Rossiter, K. A., Maceviciute, U., & Young, A. W. (2015). Personality judgments from everyday images of faces. *Frontiers in psychology*, *6*.

Talamas, S. N., Mavor, K. I., & Perrett, D. I. (2016). Blinded by beauty: Attractiveness bias and accurate perceptions of academic performance. *PloS one*, *11*(2), e0148284.

Todorov, A., Chaiken, S., & Henderson, M. D. (2002). The heuristic-systematic model of social information processing. *The persuasion handbook: Developments in theory and practice*, 195-211.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623-1626.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*, 519-545.

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological science*, *25*(7), 1404-1417.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in cognitive sciences*, *12*(12), 455-460.

Treise, D., & Weigold, M. F. (2002). Advancing Science Communication A Survey of Science Communicators. *Science Communication*, *23*(3), 310-322.

Uleman, J. S., & Kressel, L. M. (2013). A Brief History of Theory and Research on Impression Formation. In Carlston, D. E. (Ed.), *The Oxford Handbook of Social Cognition* (pp. 53-73). Oxford: Oxford University Press.

Vernon, R. J., Sutherland, C. A., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353-E3361.

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., ... & Walker-Smith, J. A. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, *351*(9103), 637-641.

Walker, M., Jiang, F., Vetter, T., & Sczesny, S. (2011). Universals and cultural differences in forming personality trait judgments from faces. *Social Psychological and Personality Science*, *2*(6), 609-617.

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, *9*(11), 12.

Walker, M., & Vetter, T. (2016). Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of personality and social psychology*, *110*(4), 609-624.

Weisbuch, M., & Ambady, N. (2011). Thin-slice vision. In *The science of social vision*. (pp. 228-247). New York: Oxford University Press.

Williams, W. M., & Ceci, S. J. (2015). National hiring experiments reveal 2: 1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences*, *112*(17), 5360-5365.

Willis, J., & Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychological science*, *17*(7), 592-598.

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology*, *67*(2), 222-232.

Wojciszke, B. (2005). Morality and competence in person-and self-perception. *European review of social psychology*, *16*(1), 155-188.

Wojciszke, B., & Klusek, B. (1996). Moral and competence-related traits in political perception. *Polish Psychological Bulletin,27*, 319-324.

Yeagley, E., Morling, B., & Nelson, M. (2007). Nonverbal zero-acquaintance accuracy of self-esteem, social dominance orientation, and satisfaction with life. *Journal of Research in Personality*, *41*(5), 1099-1106.

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and human behavior*, *15*(6), 603-623.

Zhang, Z., & Yuan, K. H. (2015). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement*, *76*(3), 387-411.

APPENDICES

Appendix A

Science engagement questions.

Strongly disagree						Strongly agree
1	2	3	4	5	6	7

1) I am knowledgeable about science

2) I find scientific ideas fascinating

3) I do not understand most scientific research

4) I like to read about scientific discoveries

5) I enjoy watching and listening to people describe scientific ideas

6) I have little interest in science

7) I am well-equipped to evaluate scientific arguments

8) I fully understand the scientific method

Online studies used a 1-5 scale, instead of 1-7 as illustrated above.

Appendix B

List of the 60 scientific article titles used for the “Interest in Research” studies, collected from ScienceDaily.com, with links to the article pages.

BIOLOGY

No.	Title	Link
1	What's next in diets: Chili peppers?	http://www.sciencedaily.com/releases/2015/02/150208152751.htm
2	We're all going to die; DNA strands on the end of our chromosomes hint when	http://www.sciencedaily.com/releases/2015/02/150206125342.htm
3	Another reason to drink wine: It could help you burn fat, study suggests	http://www.sciencedaily.com/releases/2015/02/150206111702.htm
4	Cow immune system inspires potential new therapies	http://www.sciencedaily.com/releases/2015/02/150206071230.htm
5	Drug-resistant bacteria lurk in subway stations, high school students discover	http://www.sciencedaily.com/releases/2015/02/150205174937.htm
6	Carnivorous mushroom reveals human immune trick: How we punch our way into cancer cells	http://www.sciencedaily.com/releases/2015/02/150205142913.htm

7	Medical marijuana for children with developmental and behavioral disorders?	http://www.sciencedaily.com/releases/2015/02/150205122733.htm
8	Acute psychological stress reduces ability to withstand physical pain	http://www.sciencedaily.com/releases/2015/02/150205111806.htm
9	Opinions on vaccinations heavily influenced by online comments	http://www.sciencedaily.com/releases/2015/02/150205095239.htm
10	Compound found in grapes, red wine may help prevent memory loss	http://www.sciencedaily.com/releases/2015/02/150204184230.htm
11	Possible use of medical marijuana for depression	http://www.sciencedaily.com/releases/2015/02/150204163012.htm
12	The brain's social network: Nerve cells interact like friends on Facebook	http://www.sciencedaily.com/releases/2015/02/150204134127.htm
13	Artificially intelligent robot scientist 'Eve' could boost search for new drugs	http://www.sciencedaily.com/releases/2015/02/150203204453.htm
14	Add nature, art and religion to life's best anti-inflammatories	http://www.sciencedaily.com/releases/2015/02/150203133237.htm
15	If Facebook use causes envy, depression could follow	http://www.sciencedaily.com/releases/2015/02/150203123415.htm
16	Income, education affect calorie menu use: Most notice it, few use it	http://www.sciencedaily.com/releases/2015/02/150203111914.htm
17	Simple strategies used by parents lead to	http://www.sciencedaily.com/releases/

	improvements in one-year-olds at risk for autism spectrum disorder	2015/02/150203102913.htm
18	Risk for autism increases for abandoned children placed in institutions	http://www.sciencedaily.com/releases/2015/02/150202123714.htm
19	Behaviors, preferences of picky eaters described	http://www.sciencedaily.com/releases/2015/02/150202123536.htm
20	Cognitive behavioral therapy for insomnia reduces suicidal thoughts in veterans	http://www.sciencedaily.com/releases/2015/02/150202114632.htm
21	Illusion aids understanding of autism	http://www.sciencedaily.com/releases/2015/02/150202105743.htm
22	Confidence in government linked to willingness to vaccinate	http://www.sciencedaily.com/releases/2015/02/150202105515.htm
23	Reasons why winter gives flu a leg up could be key to prevention	http://www.sciencedaily.com/releases/2015/02/150202105403.htm
24	Stress balls, DVDs and conversation ease pain, anxiety during surgery	http://www.sciencedaily.com/releases/2015/01/150130211357.htm
25	Stress shared by same-sex couples can have unique health impacts	http://www.sciencedaily.com/releases/2015/01/150130132849.htm
26	Mobile and interactive media use by young children: The good, the bad and the unknown	http://www.sciencedaily.com/releases/2015/01/150130102616.htm

27	Tweeting about sexism could improve a woman's well-being	http://www.sciencedaily.com/releases/2015/01/150130081803.htm
28	Texting may be more suitable than apps in treatment of mental illness	http://www.sciencedaily.com/releases/2015/01/150129141115.htm
29	Brain circuit that controls compulsive overeating and sugar addiction discovered	http://www.sciencedaily.com/releases/2015/01/150129125459.htm
30	Elementary teachers' depression symptoms related to students' learning	http://www.sciencedaily.com/releases/2015/02/150211084106.htm

PHYSICS

No.	Title	Link
1	How oxygen is like kryptonite to titanium	http://www.sciencedaily.com/releases/2015/02/150205142919.htm
2	The laser pulse that gets shorter all by itself	http://www.sciencedaily.com/releases/2015/01/150127111033.htm
3	Entanglement on a chip: Breakthrough promises secure communications and faster computers	http://www.sciencedaily.com/releases/2015/01/150126095707.htm
4	Exotic, gigantic molecules fit inside each	http://www.sciencedaily.com/releases/

	other like Russian nesting dolls	2015/01/150122132730.htm
5	Atoms can be in two places at the same time	http://www.sciencedaily.com/releases/2015/01/150120085919.htm
6	Shedding light on why blue LEDS are so tricky to make	http://www.sciencedaily.com/releases/2015/01/150107123936.htm
7	Doing more with less: Steering a quantum path to improved internet security	http://www.sciencedaily.com/releases/2015/01/150107082223.htm
8	'Iron Sun' is not a rock band, but a key to how stars transmit energy	http://www.sciencedaily.com/releases/2015/01/150106121507.htm
9	How electrons split: New evidence of exotic behaviors	http://www.sciencedaily.com/releases/2014/12/141223114227.htm
10	Hunt for Big Bang particles offering clues to the origin of the universe	http://www.sciencedaily.com/releases/2014/12/141223113821.htm
11	New technique could harvest more of the sun's energy	http://www.sciencedaily.com/releases/2014/12/141209101855.htm
12	New technique offers spray-on solar power	http://www.sciencedaily.com/releases/2014/12/141205124349.htm
13	Laser sniffs out toxic gases from afar: System can ID chemicals in atmosphere from a kilometer away	http://www.sciencedaily.com/releases/2014/12/141203161132.htm
14	Atomic timekeeping, on the go: New	http://www.sciencedaily.com/releases/

	approach may enable more stable and accurate portable atomic clocks	2014/11/141112203349.htm
15	How to make mobile batteries last longer by controlling energy flows at nano-level	http://www.sciencedaily.com/releases/2014/11/141106082626.htm
16	Universe may face a darker future: Is dark matter being swallowed up by dark energy?	http://www.sciencedaily.com/releases/2014/10/141031082021.htm
17	A 'Star Wars' laser bullet -- this is what it really looks like	http://www.sciencedaily.com/releases/2014/10/141022103556.htm
18	Aircraft safety: New imaging technique could detect acoustically 'invisible' cracks	http://www.sciencedaily.com/releases/2014/10/141006114055.htm
19	New technology may lead to prolonged power in mobile devices	http://www.sciencedaily.com/releases/2014/09/140926112052.htm
20	Graphene: When a doughnut becomes an apple	http://www.sciencedaily.com/releases/2014/09/140923085931.htm
21	Electronics that need very little energy? Nanotechnology used to help cool electrons with no external sources	http://www.sciencedaily.com/releases/2014/09/140910132534.htm
22	'Solid' light could compute previously unsolvable problems	http://www.sciencedaily.com/releases/2014/09/140909130810.htm
23	Why some liquids are 'fragile' and others	http://www.sciencedaily.com/releases/

	are 'strong'	2014/08/140827163447.htm
24	A centimeter of time: Cool clocks pave the way to new measurements of Earth	http://www.sciencedaily.com/releases/2015/02/150209113042.htm
25	Do we live in a 2-D hologram? Experiment will test the nature of the universe	http://www.sciencedaily.com/releases/2014/08/140826121052.htm
26	Laser pulse turns glass into a metal: New effect could be used for ultra-fast logical switches	http://www.sciencedaily.com/releases/2014/08/140826100808.htm
27	Laser device may end pin pricks, improve quality of life for diabetics	http://www.sciencedaily.com/releases/2014/08/140821141610.htm
28	Neutrino detectors could help detect nuclear weapons	http://www.sciencedaily.com/releases/2014/08/140812121644.htm
29	Grass really is greener on TV, computer screens, thanks to quantum dots	http://www.sciencedaily.com/releases/2014/08/140808110028.htm
30	Superman's solar-powered feats break a fundamental law of physics	http://www.sciencedaily.com/releases/2014/07/140730093837.htm

Appendix C

List of the full 20 re-written scientific articles used for the "Quality of Research" studies.

Physics

Don't flip out: Earth's magnetic poles aren't about to switch

https://www.sciencenews.org/article/don%E2%80%99t-flip-out-earth%E2%80%99s-magnetic-poles-aren%E2%80%99t-about-switch?mode=topic&context=43&tgt=nr

The planet's magnetic field is about 10 percent weaker today than when physicists began keeping tabs on it in the 1800s. In the geologic past, such weakening preceded geomagnetic reversals —swaps of the north and south magnetic poles. Such reversals temporarily make the planet more vulnerable to charged particles blasted off the sun that can disrupt power grids and disable satellites. Despite these factors, my recent research suggests that the Earth is not heading toward a doomsday reversal of its magnetic field.

Our study suggests that, while weakening, Earth's magnetic field is still strong by historical standards. My research team and I have retraced the strength of Earth's magnetic field over the last 5 million years, and discovered that the field has been much weaker in the past than previously thought. It appears that Earth's magnetic field is just returning back to its long-term average, not weakening toward a reversal. To determine this, we have been using lava grains, which are permanent record keepers of the magnetic field strength at the time of the eruption. However, decoding that magnetic record can be tricky. We are currently working on a better decoding technique, to support our hypothesis about the geomagnetic reversal.

Final chapter to be published, in decades-long Gravity Probe B project

https://www.sciencenews.org/article/final-chapter-published-decades-long-gravity-probe-b-project?mode=topic&context=43&tgt=nr

Gravity Probe B is an ambitious project, designed to confirm Einstein's prediction that the Earth dents and whips up the space-time around it. My research team and I have been working on the final chapter of this project, which involves investigating two phenomena: the geodetic effect (how much the Earth, and by extension, all objects with mass, warps space-time), and the frame-dragging effect (the spinning Earth should yank and twist the surrounding space-time). Under Newton's law, the axis of a gyroscope totally isolated from external forces would point in the same direction forever. But because of the geodetic and frame-dragging effects, general relativity predicts that Earth should reorient a gyroscope's axis ever so slightly.

A satellite was launched, and gyroscopes made from almost perfect spheres of quartz were used to measure two predicted effects of Einstein's general theory of relativity. Unfortunately, eliminating outside forces is a difficult task, even in space. We noticed that the ping-pong ball–sized gyroscopes were wobbling in unexpected ways, and the axis would sometimes shift and point in a new direction. We are still uncertain of what is causing these deviations, which are hundreds of times larger than the gravity-driven effects we were expecting. We are currently examining five years of data, measuring electron interactions, and hoping to verify the geodetic and frame-dragging effects.

Physicists: We Know How to Turn Light Into Matter

http://www.newser.com/story/187076/physicists-we-know-how-to-turn-light-into-matter.html

My research team and I are on the brink of turning light into matter—a process first theorized in 1934 but then described by the very men behind the idea as "hopeless to try."

The subatomic particles we've figured out how to produce will not be visible to the naked eye, but will be one of the purest demonstrations of E=mc2, Einstein's famous equation describing the "interchangeable" relationship between mass and energy. We think we can attempt the feat in the next 12 months using existing technologies. Here is what we will do in general: Two particles of light (photons) will be smashed together to create an electron-positron pair, known as a Breit-Wheeler pair in honour of the earlier researchers, Gregory Breit and John Wheeler.

The specifics of the process are as follows: First a stream of electrons will be fired into a slab of gold, creating a high-energy photon beam; then, a high-energy laser will be fired into a gold can called a "hohlraum", creating light akin to what stars emit. The first beam will then be directed into the centre of the can, and the two photon sources will collide. We are expecting electrons and positrons to come out of the can. We think it is a very clean experiment: pure light goes in, pure matter comes out. If successful, this experiment would be the first demonstration of this.

Sneezes spray 'sheets, bags and strings' of fluid

http://www.bbc.co.uk/news/science-environment-34899677

My research has mapped out, for the first time, a sequence of shifting shapes found in the fluid we eject when we sneeze. We used high-speed video footage to discover precisely how the stream of mucus and saliva breaks up into drops. It moves in sheets, bursts, bags and beaded strings during this progression.

The process is important to understand because it determines the various sizes of the final droplets - a critical factor in how a sneeze spreads germs. Modelling and helping to control that spread is the ultimate aim of my research. Several other studies had measured the size of droplets produced by sneezes, but their results were variable because the first stage of

the process was poorly understood. The part that is still a big unknown is: how are these drops actually formed and what is their size distribution?

When my team and I studied the videos, we found ourselves looking at much more than droplets at various sizes and stages. We saw droplets, but we also saw that the break-up process continues to happen outside the respiratory tract. Even more surprisingly, we saw a process that cascades from sheets, to bag bursts, to ligaments, and then the ligaments destabilise into droplets.

This procession of shapes has been observed in the flow of liquids in some industrial situations, but was a surprise in this context. It was not clear at all that we would see that in a physiological fluid, and a physiological process like a sneeze.

Dark Matter Particles Detected Deep in Mine

http://www.newser.com/story/76463/dark-matter-particles-detected-deep-in-mine.html

My team and I have been searching for traces of dark matter—the mysterious substance believed to make up most of the universe's mass—at the bottom of an old mine in Minnesota. We now believe we may have detected dark matter for the first time. Detectors we placed half a mile underground, to shield them from cosmic rays, appear to have captured two dark matter particles.

There is a one in four chance that the result may have been caused by some other effect. The invisible, subatomic particles are the core components that hold the rest of the universe together, and could explain mysteries such as why time only moves in one direction. We plan to install more sophisticated detectors in the mine next year, in the hope of overcoming the limitations imposed by chance results. Until then, our results are still

tentative, as we are awaiting feedback and suggestions on how to improve our method of capturing dark matter.

World's Next Timekeeper: Quantum Superclock?

http://www.newser.com/story/188610/worlds-next-timekeeper-quantum-superclock.html

Sick of missing appointments by milliseconds because of inaccurate atomic clocks? My research team and I have been working on using quantum physics to create a timekeeper so accurate it could help explain some of the mysteries of time itself. The "quantum superclock" would involve multiple atomic clocks, each in its own satellite orbiting the Earth and each carrying pairs of linked particles entangled in such a way that measuring a property of one of them instantaneously determines the same property for the other—a phenomenon known as "quantum entanglement". In the case of these satellites, a central satellite would fashion its clock particles in an entangled state, then extend the entanglement to another satellite, and so on, until the quantum network is created.

The linked clock would be far more accurate than anything that exists today, allowing for precise linking of financial markets, better space navigation, and even the detection of subtle shifts in space and time. In addition to measuring time, the quantum timepiece could measure the Earth's terrain so accurately that somebody digging a tunnel under the US-Mexican border could be spotted from space. Much research remains to be done before such a “superclock” can become a reality, but we're trying to be a little bit visionary. All the building blocks have been demonstrated in principle, and we want to show what might lie ahead if all these fields merge together.

All the Universe Is … Just a Hologram?

http://www.newser.com/story/179066/all-the-universe-is-just-a-hologram.html

Prepare for a head trip: The universe may actually be a hologram and everything you see an illusion. A new line of research conducted by my research team and I could be capable of proving gravity comes from thin, vibrating strings—holograms of events in a simpler, flatter cosmos. It was an idea first put forth by physicist Juan Maldacena in 1997, but never tested until now. My colleagues and I propose mathematical evidence, via two studies on black holes, that this hologram theory may in fact be right. If proven, it would solve inconsistencies in Einstein's theory of gravity, and be a solid footing for string theory.

Our theory is similar to the security chip on your credit card: It's a 2D surface that has all the data needed to describe a 3D object. Basically, all the information about our universe is stored in a flattened version that projects everything we see. We think this is rather curious. Our new research shows that the thermodynamics of certain black holes can be reproduced from a lower-dimensional universe, but the universes we explored do not look like our own. Still, it does show hope that our universe can be explained by a similar theory. We invite our readers to make of the findings what they will, while we are working on addressing the inconsistencies and unexpected results of our project.

Hottest Temp Ever Created by a Human: 7 Trillion Degrees

http://www.newser.com/story/149095/hottest-temp-ever-created-by-man-7-trillion-degrees.html

My research team and I have recently managed to smash gold ions into a quark-gluon plasma much like the one believed to have existed in the milliseconds after the Big Bang. The plasma hit a mind-boggling 7.2 trillion degrees Fahrenheit, making it 250,000 times hotter than the core of the sun—and in the process broke the Guinness World Record for the hottest man-made temperature ever recorded.

During our study, we ran gold ions in both directions through a particle accelerator, smashing them together at speeds so fast that the neutrons and protons in their nuclei melted. This produced a nearly frictionless fluid, which, interestingly, has also been observed in atoms near absolute zero—that is 10 million trillion times colder than the quark-gluon plasma we created. We expected to reach these temperatures, but we did not at all anticipate the nearly perfect liquid behaviour. We are currently working on understanding why the behaviour we observed occurred and we are awaiting feedback on our work from experts in the area.

Lisa Pathfinder launches to test space 'ripples' technology

http://www.bbc.co.uk/news/science-environment-34985807

Recently, my team of researchers and I have launched the Lisa Pathfinder satellite, designed to test the technologies needed to detect gravitational waves. These are a prediction of Einstein's Theory of General Relativity, and describe the warping of space-time produced by cataclysmic events in the cosmos. Having such a capability would make it possible to detect the merger of monster black holes - a marker for the growth of galaxies through time.

Pathfinder contains just a single instrument, which is designed to measure and maintain a 38cm separation between two small gold-platinum blocks. These will be allowed to free-fall inside the spacecraft, and a laser system will then monitor their behaviour, looking for path deviations as small as a few picometres. The signal, though, is expected to be extremely subtle –while this precision performance is relatively routine in Earth laboratories, it is very exacting to try to demonstrate it in space.

We are particularly keen to study supermassive black holes because their creation and evolution seems to be tied inextricably to that of the galaxies that host them, and probing their properties would therefore reveal details about how the great structures we see on

the sky took shape through cosmic history. Although we have no results as of yet, we are hoping to collect some revealing data in the near future.

Bloodhound Diary: It's rocket science

http://www.bbc.co.uk/news/science-environment-35011505

My research team and I have been working on developing a car that will be capable of reaching 1,000mph (1,610km/h). Powered by a rocket bolted to a Eurofighter-Typhoon jet engine, the vehicle will first mount an assault on the world land speed record (763mph; 1,228km/h).

We need some form of rocket system in order to reach 1000+ mph, as jet engines alone won't be enough - after all, we're trying to go faster than any jet fighter has ever been at ground level, so we're above the design speed of any known jet engine. Solid rockets (like very large fireworks) can't easily be controlled or shut down, so they are not a favourite of mine. Liquid rockets (the sort used for "normal" space rocket launches) work by mixing two very excitable liquids together and trying to control the very angry reaction it causes. Liquid rockets are very powerful, but the liquids are not nice to use (or to carry in large quantities in the car) so once again this is not ideal.

Hence our choice was for a hybrid rocket system. The solid fuel "grain" is made from a synthetic rubber, while concentrated hydrogen peroxide, known as "high-test" peroxide (or HTP for short) gives us a fairly well-behaved oxidiser. These make for a safe payload in rocketry terms. These chemicals are certainly a lot friendlier than liquid hydrogen, liquid oxygen, various solid fuel "explosives", etc., that other rockets use. At the moment, the rocket fuel is still unstable, so we are performing further tests before launching the rocket.

Biology

DNA editing shows success in mosquito sterilization

https://www.sciencenews.org/article/dna-editing-shows-success-mosquito-sterilization?mode=topic&context=87&tgt=nr

A new genetic tool may help eradicate Africa's main malaria-carrying mosquitoes. My colleagues and I have been working on a self-propagating cut-and-paste system, known as a gene drive, which could sterilise female mosquitoes. Instead of stopping the mosquitoes from transmitting the parasite, this new gene drive aims to eliminate the mosquitoes themselves by making it impossible for females to reproduce.

Gene drives are engineered pieces of DNA designed to slice a target gene and insert themselves. Like Star Trek's Borg, gene drives assimilate every unaltered target gene they encounter. These ambitious bits of DNA break standard inheritance rules to get passed on to more than 50 percent of an altered animal's offspring, "driving" themselves quickly through populations. We are planning to combine approaches, by first releasing a gene drive that would prevent mosquitoes from carrying malaria, then later releasing one to control mosquito populations.

The latest mosquito gene drive works in laboratory settings, but further modifications may be needed before it is ready for release in the field. This gene drive had some technical glitches, so it won't be the final version that we would release to control wild mosquito populations – we are currently working on improving the gene drive. However, we are hopeful that future gene drives could curb populations of mosquitoes.

Porpoises Use Sound Like a Flashlight

http://www.newser.com/story/204800/porpoises-use-sound-like-a-flashlight.html

In a recent study, my colleagues and I found that porpoises, which are aquatic marine mammals, have the amazing ability to not only locate prey with a beam of sound, but adjust the field of clicks and buzzes as they move in for the kill, preventing the fish from slipping away. The switch, as we discovered, is much like adjusting a flashlight. Imagine you're looking for your car in a parking lot. You could use a narrow beam over a long distance and still see a lot, but when you're trying to get your keys into the car, you would switch to a wider beam. This is similar to what we see in porpoises. The beam is controlled by a fatty structure in the porpoise's head, called the melon.

Our recent research suggests that like some bats, harbour porpoises can broaden their bio-sonar beam during the terminal phase of attack but, unlike bats, maintain the ability to change beam-width within this phase. After studying harbour porpoises in a semi-enclosed research facility that gave the animals seafloor access, we believe that other dolphins and whales have the same sonar ability. Our discovery suggests many porpoises end up in fishing nets because of "attention blindness" that causes them to ignore potential hazards as they zero in on a fish. Our results cannot be generalised to other animals which might have similar sonar abilities, and we are currently addressing this issue.

Crocodiles May Be Watching You While They Sleep

http://www.newser.com/story/214872/crocodiles-may-be-watching-you-while-they-sleep.html

Crocodiles may be keeping an eye on humans even when the crocs are sleeping, according recent research conducted by my colleagues and I. We have been monitoring juvenile crocodiles using infrared cameras and determined they often slept with one eye open and may only sleep with half their brain at a time. This type of unihemispheric sleep has been observed in birds and aquatic mammals, such as the dolphin. These findings are the first of

their kind involving crocodilians and may change the way we consider the evolution of sleep; what we think of as 'normal' sleep may be more novel than we think.

By sleeping with one eye open and half their brain active, crocodiles could respond quickly to threats and prey—including humans. They definitely monitored the human when they were in the room. But even after the human left the room, the animal still kept its open eye … directed towards the location where the human had been. We are planning more tests using infrared cameras placed directly in the crocodiles' enclosure, to confirm our findings in crocodiles and other reptiles; specifically, more research is needed to determine whether one half of the crocodilian brain is actually awake as the other half sleeps. Depending on what we find, our version of whole-brain sleep—which we think of as normal—could actually be an evolutionary oddity.

Study Overturns Long-Held Belief on Hummingbirds

http://www.newser.com/story/211581/study-overturns-long-held-belief-on-hummingbirds.html

Hummingbirds beat their wings approximately 50 times per second, but that's nothing compared to how fast they can drink. My current research aims to debunk nearly 200 years of scientific thinking on how hummingbirds accomplish that task. Scientists have long believed hummingbirds drank by wicking, a process that allows liquid to flow through small spaces without benefit of gravity. My colleagues and I were sceptical of this slow method because it would limit hummingbirds' energy intake. Hence, we set off on a five-year study to find out how the birds really do their drinking.

After recording 18 hummingbird species drinking from specially made artificial flowers in the wilds of the US, Ecuador, Brazil, and Colombia, we believe that hummingbirds drink by using their tongues as elastic micropumps. After zipping toward a flower, the hummingbird

flattens its outstretched tongue, and the compressed tongue remains flattened until it contacts the nectar. After contact with the nectar surface, the tongue reshapes, filling entirely with nectar. The last move involves a bend of the tongue to pull in the nectar, and it all takes place in less than a tenth of a second—an impressive feat for something thinner than a fishing line.

Our results suggest the tiny birds can sip up to 10 drops of nectar every 15 milliseconds. It turns out that hummingbird tongues do not wick—they pump. However, these results still need to be verified both in a laboratory setting, and through slow-motion video recording of hummingbirds in the wild.

Study Suggests Earth Life Began on Mars

http://www.newser.com/story/173372/did-all-earth-life-begin-on-mars.html

Were our earliest ancestors Martians? A recent study conducted by my colleagues and I suggests that all life on Earth may have originated on the Red Planet. This could be because Mars would have had plenty of the minerals that are best at forging RNA (Ribonucleic Acid), which is one of the key components of life and is believed to have predated DNA. On Earth, those minerals would have dissolved into the ocean (water is corrosive to RNA). But life could have formed on Mars, then headed here on meteorites; however, we are not the first to propose such a theory.

Our investigation centres on how atoms were arranged to form RNA, DNA, and proteins. Minerals containing the elements boron and an oxidized form of molybdenum were central to the process—but at the time, Earth was probably incapable of supporting enough of such minerals. The analysis of a Martian meteorite revealed the presence of boron on Mars, and we now believe that the oxidized form of molybdenum was there, too. As such, the evidence seems to be building that we are actually all Martians; that life started on Mars

and came to Earth on a rock. It's lucky that we ended up here, nevertheless—as certainly Earth has been the better of the two planets for sustaining life. Although the presence of boron and molybdenum is encouraging, this evidence is currently insufficient to fully support our hypothesis.

Things Can Actually Live at the Ocean's Deepest Point

http://www.newser.com/story/164587/things-can-actually-live-at-the-oceans-deepest-point.html

The Pacific Ocean's Mariana Trench contains the deepest point in all the world's oceans. But despite its nearly eight-mile depth (Mount Everest, by comparison, does not hit six miles),the lowest point of Mariana Trench – also known as Challenger Deep -- is also home to life, my recent research suggests. My research team and I sent a robot into Challenger Deep to measure the oxygen being consumed at the spot, an indicator of life. An analysis of the sediment it recovered there points to the presence of 10 times the amount of bacteria it identified at an area only about half that deep nearby.

Video footage taken on the floor also showed that some far bigger creatures live down there: our team was able to trap crustaceans, known as Hirondellea gigas, which measure less than an inch in length. It's all a surprise because it is so dark down there—and most underwater food chains are reliant on photosynthetic plankton that need light. And while organic matter from the surface does filter downward, only 1% to 2% of it makes it to the ocean's average depth of 2.3 miles. We believe the trench catches nutrients when earthquakes rattle the surrounding area. It acts as a trap just because it's a big hole; our findings point to the possibility of thriving life in other trenches. We are currently working on providing more evidence for our findings, by conducting a more in-depth chemical analysis.

Search Begins for Life in Antarctic Lake

http://www.newser.com/story/159187/search-begins-for-life-in-antarctic-lake.html

What lurks in the pitch black, near-freezing waters of Lake Ellsworth? That is what my team of researchers and I hope to find out soon, after we have begun our trek to the lake. We have begun drilling through more than two miles of ice to reach the water, kept just above freezing by the rocks beneath it. And while others have drilled into Antarctic lakes before, this marks the first time it will be done using ultra-sterilized equipment. Unless we keep the experiment very clean, we're likely just to measure the things that we bring down us with, which would be pointless.

It will take us five days to create the borehole (the deepest ever made in this fashion) using a high-pressure hose that blasts sterilized water heated to about 194 degrees Fahrenheit. We will then have to rapidly collect samples before it freezes closed—and we hope to find microbial life that has developed in ways never seen before. Should we succeed in doing so, it could have out-of-this-world implications. If there's life on Jupiter's Europa it'll be living in a very similar way to life in Lake Ellsworth, with total darkness, lots of pressure, and using chemical processes rather than sunlight to power biological processes. We are hoping for results early next week, although we are concerned about the implications of contaminating the samples.

Beneath Pacific Lies Ancient, Barely Alive Bacteria

http://www.newser.com/story/146273/beneath-pacific-lies-ancient-barely-alive-bacteria.html

Some 100 feet below the most nutrient-starved part of the Pacific Ocean floor, incredibly old life exists. In the most detailed look yet at the lifestyles of "extremophile" bacteria, my

research team and I are looking into the possibility that the organisms have survived for what could be as long as millions of years solely on whatever nutrients were around when the sediment settled around them. These communities have not received input or new food since the dinosaurs walked the planet. The communities that are left down there are the ones that can deal with the lowest amount of food.

The metabolisms of the deep-sea bacteria are incredibly slow. So far we have found it impossible to determine whether they reproduce—which could likely happen only once every few thousand years at the fastest—or are many millions of years old, having repaired themselves over the eons. These organisms live so slowly that when we look at it at our own time scale, it's like suspended animation. We are currently working on devising a new method of determining the bacteria's lifestyle -- creating a computer simulation of their evolution, based on a time-lapse video of the bacteria -- and we are awaiting support and suggestions from colleagues. The main lesson here is that we need to stop looking at life at our own time scale.

Slime Mould Is Smarter Than You Think

http://www.newser.com/story/136416/slime-mold-is-smarter-than-you-think.html

It may not look like much, but slime mould is capable of human-like "thought" beyond the reach of the most sophisticated computers. The organism can arrange its cells in order to find the quickest route through a maze, as our most recent research suggests. Humans are not the only living things with information-processing abilities; simple creatures can solve certain kinds of difficult puzzles. If you want to spotlight the essence of life or intelligence, it's easier to use these simple creatures.

This is why slime mould could be the key to building the computers of the future. The mould can recall stressful situations and has also been seen to form itself into patterns reminiscent

of sophisticated railway systems. Computers are not so good at analysing the best routes that connect many base points because the volume of calculations becomes too large for them. We have measured the time taken and number of errors made by slime moulds when escaping a maze and it seems that slime moulds, without calculating all the possible options, can flow over areas in an impromptu manner and gradually find the best routes. Although we have no concrete evidence supporting our hypothesis as of yet, we are currently working on computer-model based on mould, and we hope to receive feedback on this model shortly.

Earth Holds 8.7M Species, and Most of Them are Still Undiscovered

http://www.newser.com/story/126729/lalapalooza-earth-holds-87m-species-study-finds.html

Humanity shares the planet with roughly 8.7 million species, most of them still undiscovered, as our new line of research is suggesting. To illustrate this, we are comparing the planet to a machine with 8.7 million parts, each performing an important function. If you think of the planet as a life-support system for our species, you want to look at how complex that life-support system is. We're tinkering with that machine because we're throwing out parts all the time. The research shows that we are really fairly ignorant of the complexity and colourfulness of this amazing planet. We need to expose more people to those wonders; it really makes you feel differently about this place we inhabit.

My research team and I have used complex mathematical models to tackle a question that has long puzzled scientists, identifying numerical patterns in data from 1.2 million known species, excluding viruses and microorganisms. Previous estimates ranged from 3 million to 100 million species. We believe that an astonishing 86% of terrestrial species and 91% of marine species are still unknown. However, these results are merely mathematical

estimates, and not the result of a thorough investigation. We will be unable to provide an accurate number until all species are discovered and counted.