

# Reproducible research into human chemical communication by cues and pheromones: learning from psychology's renaissance

Tristram D. Wyatt<sup>1,2</sup>

[Version accepted 16 January 2020] DOI: 10.1098/rstb.2019-0262 References in Harvard format.

In a themed issue of *Philosophical Transactions Royal Society B* with provisional title "Advances in human chemosignaling" to be published in 2020, containing the contributions from a Royal Society Discussion Meeting held 1-2 April 2019 entitled [Chemical Communication in humans](#)

<sup>1</sup>Department of Zoology, University of Oxford, 11a Mansfield Road, Oxford, OX1 3SZ UK

and

<sup>2</sup>Centre for Biodiversity and Environment Research, University College London, Gower Street, WC1E 6BT, UK.

Correspondence:

Tristram D. Wyatt, Department of Zoology University of Oxford, 11a Mansfield Road, Oxford, OX1 3SZ, UK

tristram.wyatt@zoo.ox.ac.uk

ORCID <http://orcid.org/0000-0003-0371-0713>

Cite as Wyatt, TD (2020 [in press]) Reproducible research into human chemical communication by cues and pheromones: learning from psychology's renaissance. *Phil Trans B* DOI: 10.1098/rstb.2019-0262

Running head: Human chemical communication research: learning from psychology

11,695 words, 1 figure, 1 box, 1 supplementary data file

Key words:

Reproducibility, Registered Reports, olfaction, false positive, pre-registration, semiochemical

## Summary

Despite the lack of evidence that the “putative human pheromones” androstadienone and estratetraenol ever were pheromones, some 60 studies have claimed ‘significant’ results. These are quite possibly false positives and can be best seen as potential examples of the ‘reproducibility crisis’, sadly common in the rest of the life and biomedical sciences, which has many instances of whole fields based on false positives. Experiments on the effects of olfactory cues on human behaviour are also at risk of false positives because they look for subtle effects but use small sample sizes. Research on human chemical communication, much of it falling within psychology, would benefit from vigorously adopting the proposals made by psychologists to enable better, more reliable science, with an emphasis on enhancing reproducibility. A key change is the adoption of study pre-registration and/or Registered Reports which will also reduce publication bias. As we are mammals, and chemical communication is important to other mammals, it is likely that chemical cues are important in our behaviour and that humans may have pheromones, but new approaches will be needed to reliably demonstrate them.

181 words

## 1. Introduction

The study of human chemical communication, covering responses to human olfactory cues and possible human pheromones, is potentially at a turning point together with the rest of psychology. Even if we do not always think of research into human chemical communication as a branch of psychology, it is, as much of it involves human behavioural responses. Our field is thus likely to be a victim of the ‘reproducibility crisis’ in psychology and other life sciences but it also could be a beneficiary of the ideas rapidly emerging from psychology to improve the reproducibility and reliability of experiments (Bishop 2019; Chambers 2017; Munafò *et al.* 2017; Nelson *et al.* 2018).

Why are new ways of working needed? As I will explore in this paper, there are good reasons to doubt the most studied ‘putative human pheromones’, androstadienone and estratetraenol, despite their popularity with experimenters (Doty 2010, 2014; Wyatt 2015; Wysocki & Preti 2004). Some 60 papers report close to uniformly positive results from tests with these molecules (Supplementary File 1), but it is quite possible that these are false positives. This might be surprising to some readers, but it is consistent with patterns demonstrated in other areas of the life sciences which have entire fields based on founding studies which turn out to be unreplicable false positives (spurious ‘significant’ results which cannot be replicated) (Section §5). For example around 450 papers built on the now disproved link between an unusual version of the serotonin transporter gene 5-HTTLPR and depression (Alexander 2019).

Similarly, much of the literature on human chemical communication exploring cues, such as odours of familiar and unfamiliar people, reports a search for subtle effects. However, we tend to use small-scale experiments with small sample sizes: across the rest of psychology, typically only 50-70 % of small-scale experiments can be replicated (Section §5).

I will start by introducing different kinds of chemical communication (semiochemicals) (Section §2), defining cues and pheromones, outline the essential steps for identifying a pheromone, and an operational definition. I then summarize the evidence that androstadienone and estratetraenol are highly unlikely to be human pheromones (Section §3). After explaining briefly why human chemical

communication research largely comes within psychology (Section §4), I explore the evidence for a ‘reproducibility crisis’ in psychology and the rest of the life sciences (Section §5) and the constructive responses to reduce the problem, ideas from what has been called ‘psychology’s renaissance’ (Section §6) (Bishop 2020; Chambers 2017; Munafò *et al.* 2017; Nelson *et al.* 2018). In ‘A manifesto for reproducible science’ aimed at researchers in human behaviour, Munafò *et al.* (2017) present ways to reduce bias in data collection and analysis: one key proposal is study pre-registration and/or ‘Registered Reports’ explained in Section §7. A recently debunked ‘social priming’ effect, ‘power posing’, illustrates both the problem and how study pre-registration helped resolve it. Finally, I briefly discuss positive proposals for how we can take forward studies on human chemical communication, notwithstanding the many challenges this task will present (Section §8).

I should say at the outset that I offer these observations respectfully: I am not a scientist actively experimenting in human chemical communication, but I can perhaps offer a disinterested perspective.

## 2. Chemical communication: cues and pheromones

Chemical communication in humans involves responses to odour cues and, potentially, pheromones (evolved chemical signals, which I will define shortly). Odour cues can be used as information by the receiving individual but did not evolve for this function. For example, a mosquito uses carbon dioxide emitted by its mammal host as a cue to locate it. The mammal does not release carbon dioxide in its breath as an evolved signal to attract mosquitos, but the insect has evolved to respond to this stimulus (Cardé 2015).

### (a) Cues to health, individual identity, mate choice, and physiological state

Humans may be able to use information from odour cues in a variety of ways, such as detecting and avoiding infected individuals (Olsson this issue). Different infections and other diseases produce their own characteristic smells (Shirasu & Touhara 2011). One possibility is that animals, including humans, have evolved specific responses to these smells of infection. More likely is that the response is a more general one: to detect and avoid a conspecific that is not ‘smelling right’, as compared with uninfected conspecifics (or self). This would be more generalizable for any new infection not previously met in the population.

Our individual odours may influence mate choice. The highly variable chemical profile which differs between individual people can be learnt as an odour ‘fingerprint’ (see references in Wyatt 2010, 2014). This memory can be used to recognise siblings, neighbours, or partners. The memory of the odours of siblings when growing up together may be used as a cue to avoid these kin as mates when adult. Proteins of the Major histocompatibility complex, MHC (also known as HLA, human leukocyte antigen), a key part of the immune system, contribute to individual differences in odour profile, by mechanisms still not fully understood. The evidence for MHC influences on mate choice in humans is still debated (Havlíček *et al.* this issue; Lobmaier *et al.* 2018).

Physiological changes in your body may be reflected in the odour molecules you give off, providing cues for others to gauge your internal state. Experiments suggest that the odours of fear or stress may be detected by other humans as cues (de Groot this issue). Human males may be able to detect changing odours over women’s menstrual cycles (Haselton & Gildersleeve 2016; Havlíček *et al.* 2006).

## (b) Pheromones

Pheromones are molecules that are evolved signals, in defined ratios in the case of multiple component pheromones, which are emitted by an individual and received by a second individual of the same species, in which they cause a specific reaction, for example, a stereotyped behaviour or a developmental process ((Wyatt 2010), modified after (Karlson & Lüscher 1959)). Pheromones are chemical signals which have evolved with this function in the signaller (Maynard Smith & Harper 2003; Wyatt 2014, 2017).

Pheromones are the same in all sexually mature males of a species, for example. It is the consistency in these molecules between individual males in a population which allows them to be identified as a pheromone. Some males may produce more of the pheromone and thus may be more attractive to females: for example, well-fed male voles with high testosterone levels produce more pheromone (Ferkin *et al.* 1994). Many pheromones, perhaps most, consist of a particular combination of molecules not a single molecule (Wyatt 2014, 2017).

The steps needed to establish that a molecule(s) is a pheromone start with describing a behavioural or physiological response which is mediated by a potential chemical stimulus such as a secretion. The experimenter then develops a repeatable experiment called a bioassay which allows the response to the odour to be measured. The bioassay allows the experimenter to track activity as samples are collected and analyzed. Having identified, and synthesized the candidate bioactive molecule(s), the final task is to confirm that the proposed molecule(s), at natural concentrations, are necessary and sufficient to recreate the response with the original bioassay (Wyatt 2014, p49 ff). These steps or their equivalent, including rigorous bioassays, remain an essential part of pheromone identification today. To be credible, any claim that a molecule or combination of molecules is a pheromone must include the publication in full of this systematic approach (Wyatt 2017).

I have proposed an operational definition of pheromones to specify the minimum evidence needed to pragmatically demonstrate a pheromone (Box 1) (Wyatt 2014, 2017). Among the features distinguishing cues and pheromones, the formal definition of a pheromone specifies that production has evolved for that function (Maynard Smith & Harper 2003, p. 3). This criterion is easily satisfied in female sex pheromones in moths for example, where we understand the detailed genetics of the production and release of these well characterized pheromones (Wyatt 2014). However, for many otherwise respectable pheromones, we do not know enough about the ways in which production and/or reception may have evolved. So, I have proposed that we formalize an operational definition of pheromone, which most people already use in practice, as ‘fully identified molecule(s), the same across a species, in all lactating mature females for example, which when synthesized elicit the same characteristic response in the conspecific receiver as the natural stimulus.’ To legitimately assert that a molecule or specific combination of molecules qualifies as a pheromone for a species it would need to satisfy the criteria summarized in Box 1. There are many mammal pheromones which satisfy the criteria (Ishii & Touhara 2018; Liberles 2014; Tirindelli *et al.* 2009) (despite the doubts expressed in Doty (Doty 2010), addressed by Wyatt (Wyatt 2014, 2017)). However, the molecules proposed as human pheromones fail to meet the criteria (next section).

Box 1: Pheromones: an operational definition. Adapted from Wyatt (2014) with permission.

1. **The synthesized molecule/combination of molecules should elicit the same response as the natural stimulus in the bioassay.** This is the fundamental basis for the designation of a pheromone.
2. **It should act in this way at natural concentrations.** Concentration is important for mammals and other animals. At high concentrations, spurious results may occur as non-pheromones may stimulate receptors.
3. **For multicomponent pheromones, experiments should demonstrate that all compounds in the combination are necessary and sufficient to elicit the full response.**
4. **Only this molecule or the proposed combination of molecules elicits the effect** (unlike other similar molecules or combinations that the animal would normally encounter).
5. **There should be a credible pathway for the pheromone signal to have evolved by direct or kin selection.** In evolutionary terms, to be a signal, both the emission and reception of the pheromone signal should have evolved for a particular function.

### 3. 'Putative human pheromones'

Interest in 'putative human pheromones' has been focussed on two main areas, menstrual synchrony and sexual attraction.

#### (a) 'Menstrual synchrony'

A pioneering study by McClintock (1971) reported the convergence of menstrual cycles of women living in close proximity. A pheromonal basis for the effect, menstrual synchrony, was proposed by Stern and McClintock (1998) who placed extracts of armpit secretions from women at different stages in their cycles on the upper lip of other women. However, ever since McClintock (1971) there has been lively debate about whether the phenomenon of synchrony itself really exists: while some studies have found menstrual synchrony, many other studies have failed to do so. Methodological questions and statistical doubts suggest that the phenomenon of synchrony might be an artefact (for references see Doty 2010, p. 168 ff; 2014). To my knowledge, no follow-up to Stern and McClintock (1998) has been published and no molecules have been proposed as pheromones. My feeling at this point is that menstrual synchrony is not supported by the evidence.

#### (b) Sexual attraction

There have been three waves of 'putative human pheromones' for sexual attraction, first with 'copulins' in the 1970s, second with androstenone and androstenol in the 1980-1990s, and third, from 1991, the steroid molecules androstadienone and estratetraenol (reviewed and critiqued in Doty 2014; Havlíček *et al.* 2010; Wyatt 2015; Wysocki & Preti 2004). In this paper I will focus on the 'third wave', as androstadienone and estratetraenol are the molecules still currently widely studied.

The literature on androstadienone and estratetraenol starts with the 1991 paper by Monti-Bloch and Grosser (1991) in a conference proceedings sponsored by the EROX Corporation (which was patenting the two molecules as 'putative human pheromones') (Wyatt 2015). As detailed in Wyatt (2015), no information was given in Monti-Bloch and Grosser (1991), nor in any patents, about how

these molecules were found and shown to be pheromones – we don't even know which parts of the body the original samples came from.

The Monti-Bloch and Grosser (1991) paper was not one that *could* have established that the proposed molecules were pheromones. It was not designed to. It basically only reported the test of 5 molecules, all supplied without further justification by the EROX Corporation. The samples were only tested on 20 men and 20 women. A further problem is that the recordings were claimed to be from the 'Vomeronasal organ' (VNO) but subsequent research has concluded that human adults, like other great apes and some other mammals, do not have a functioning VNO (Smith *et al.* 2014). There is no published paper or combination of studies anywhere which provides any evidence, meeting the defining steps outlined in Section §2b, that androstadienone and estratetraenol are pheromones. (Studies that simply take it for granted that these are human pheromones cannot be taken as evidence).

There is no more reason to use these two molecules than any other pair of molecules chosen at random from the thousands of molecules emitted from the human body. There is no more reason now than in the year 2000 when Jacob and McClintock (2000) reported a now widely cited study using the molecules (used only because of Monti-Bloch and Grosser (1991)'s claims).

In the almost 30 years since 1991, there have been 60 published experimental studies using androstadienone and/or estratetraenol, most of them since 2000 (Web of Science) [Supplementary File 1]. The popularity of these molecules shows no signs of diminishing: 20 of these studies were published in the 5 years 2015-2019, including 5 in 2019. The studies have shown almost uniformly positive results rejecting the null hypothesis, though some studies have been contradictory. Only since 2015 have a few negative results started to be published (Ferdenzi *et al.* 2016; Hare *et al.* 2017). In the early 2000's, scientists could argue that a leading scientist of the time, McClintock, had endorsed these molecules as potentially 'putative human pheromones', though in 2000 Jacob and McClintock (2000) were cautious (Wyatt 2015). Recent papers seem to avoid citing the inconvenient papers which question the validity of using these molecules (Doty 2010, 2014; Wyatt 2015; Wysocki & Preti 2004), or if cited, side step the implications.

The research rationale for still using the molecules seems to be that since so many studies have reported positive results, androstadienone and/or estratetraenol must have effects. Each paper has an introduction citing such studies in justification. However, as explored in Section §5, positive results, including physiological measures, can be baseless. While we cannot rule out that these two molecules might have some (non-pheromonal) psychological/physiological effect(s), as any molecules might, why have studies almost exclusively used these two molecules alone without any evidence to justify using these rather than any other molecules emitted by humans, which by this argument might also have effects? In addition, the studies do not look for *any* effects, rather they assume that the molecules are sex-differentiated 'putative human pheromones', despite this never having been properly demonstrated. The most parsimonious explanation for the reported positive results is that these are quite possibly, if not highly likely to be, false positives, in particular in the light of how easy it is to produce 'false positives', explored in Section §5. It should also be remembered that there is nothing that suggests steroid molecules are more likely to be pheromones than any other kind of molecule (Doty 2010, 2014; Wyatt 2015).

Few scientists study human chemical communication and by focussing on androstadienone and estratetraenol too many of them are building further floors on an structure which has, at best, very shaky foundations (Doty 2010, 2014; Wyatt 2015; Wysocki & Preti 2004). We should be asking

ourselves: Would we risk the next 20-30 years of research trying any two other molecules found in human armpits (or any other secretion), chosen at random without solid evidence of effect?

#### 4. Human chemical communication research is a branch of psychology

Much of the research into human chemical communication including work on ‘human pheromones’ and olfactory cues comes within the field of psychology, whether it is psychophysics or experimental observations of people responding to odours such as armpit secretions collected from people under stress. This is true whether the work is published in journals such as *Psychological Science* or in a wide range of journals, such as *Chemical Senses* and *Hormones and Behavior*, which do not reference ‘psychology’ in the journal title. For example, here are some typical papers:

‘Putative human pheromone androstadienone attunes the mind specifically to emotional information.’ (Hummer & McClintock 2009)

‘Chemosignals communicate human emotions.’ (de Groot *et al.* 2012)

‘A putative human pheromone, androstadienone, increases cooperation between men.’ (Huoviala & Rantala 2013)

This is important because it means that the contemporary debates about how to make psychology more rigorous and reproducible can readily inform the ways we could transform the study of human chemical communication (Sections §5-§8).

#### 5. The ‘reproducibility crisis’ in psychology and other life sciences

Much has been written about psychology’s ‘reproducibility crisis’, which came to a head in 2010-2012 (Chambers 2017; Nelson *et al.* 2018; Nuzzo 2015; Yong 2012). Among the triggers was Doyen *et al.* (2012)’s failure to replicate the results of a highly cited ‘text-book’ social psychology experiment termed ‘social priming’: Bargh *et al.* (1996) had reported that unconsciously priming young people with words associated with elderly people made the young people walk more slowly.

However, the reproducibility problem is much wider than psychology. It has been shown, for example, in drug discovery, with the biotech company Amgen able to replicate only 6 out of 53 landmark studies in oncology and haematology (Begley & Ellis 2012), in translational biomedical research e.g. (Curran 2018), and in animal behaviour e.g. (Wang *et al.* 2018). Some non-reproducible pre-clinical papers have generated a whole field of study, with hundreds of secondary publications that expanded on ideas in the original study but which did not test its fundamental basis (Begley & Ellis 2012), in an echo perhaps of ‘putative human pheromones’.

I should say at this point that the terminology of replication, replicability, and reproducibility has not yet stabilized. Goodman *et al.* (2016) offer a useful discussion, in which they propose the following terms to clarify the different implied meanings (in different scientific fields) of the existing terminology:

- *Methods reproducibility*: provide sufficient detail about procedures and data so that the same procedures could be exactly repeated.
- *Results reproducibility*: obtain the same results from an independent study with procedures as closely matched to the original study as possible.
- *Inferential reproducibility*: draw the same conclusions from either an independent replication of a study or a reanalysis of the original study.

(Plesser 2018) based on (Goodman *et al.* 2016)

Were the conspicuous results-reproducibility failures of single psychological studies such as the Bargh *et al.* (1996) ‘social priming’ experiment typical of psychology as a whole? To explore the results-reproducibility of psychology more generally, a collaboration involving hundreds of scientists worked together to replicate 100 of the most important studies published across three leading psychology journals (Open Science Collaboration 2015). Only 40% of the original studies were judged to have results successfully reproduced (though see Nelson *et al.* (2018) for a suggestion that a higher proportion might be interpreted as succeeding - negative results in a replication might be a consequence of chance, meaning no effect was detected even though it is there). Whereas 97% of the original studies had significant results ( $P < 0.05$ ), only 36% had in the replications. Mean effect sizes were halved.

Reliability of results may be inversely related to journal rank (Brembs 2018), and as might be expected, replicability is no higher in ‘high impact’ journals. Camerer *et al.* (2018) evaluated the replicability of 21 social science experiments that were published in *Nature* and *Science* between 2010 and 2015. Only 66% of the original studies could be replicated (with a significant effect in the same direction as originally published). In this large consortium exercise, in many cases involving the cooperation of the original researchers, and despite much larger sample sizes in the experiments, the effect size was uniformly reduced.

These kinds of results, including another major multisite replication study (Many Labs 2) which could replicate only half of the original 28 studies (Klein *et al.* 2018), led science journalist Ed Yong (2018) to write ‘it seems that one of the most reliable findings in psychology is that only half of psychological studies can be successfully repeated’.

#### (a) Why is science going so wrong?

‘... many researchers persist in working in a way almost guaranteed not to deliver meaningful results. They ride with what I refer to as the four horsemen of the reproducibility apocalypse: publication bias, low statistical power, *P*-value hacking and HARKing (hypothesizing after results are known).’ Dorothy Bishop (2019)

Scientists, doing the best research they can, have nonetheless created a situation of unreliability. This is in large part because of a research culture which, under the career pressures of publication and grant-getting, has adopted and rewarded research practices (including the four horsemen above) now increasingly recognized as questionable (Bishop 2019; Bishop 2020; Chambers 2017; Munafò *et al.* 2017; Nelson *et al.* 2018). I will briefly summarize the problems here but for more detailed accounts see, for example, short papers by Munafò *et al.* (2017), Nelson *et al.* (2018) and Bishop (2020). For a longer, entertaining and practical discussion see the book *The Seven Deadly Sins of Psychology* by Chris Chambers (2017).

The first ‘horseman’ is publication bias. Novel, ‘statistically significant’, exciting, and seemingly ‘clean’ results with a ‘good story’ are more likely to be published, especially in the most competitive ‘high impact’ journals. We internalise this, so we only submit such papers (the unsubmitted studies stay unloved in the ‘file drawer’, though see below). The result in psychology, and much of the life sciences, is a literature which reports positive results in 95% of published papers – the null hypothesis is almost always rejected (Chambers 2017; Fanelli 2010; Munafò *et al.* 2017).



The second is small-scale experiments with low statistical power, with little chance of finding a nominally statistically significant finding that actually reflects a true effect and that when a true effect *is* discovered gives an exaggerated estimate (Button *et al.* 2013). Instead there is a high chance of false positives – much much higher than the ‘1 in 20’ that a  $p < 0.05$  significance threshold suggests, especially when combined with flexible statistical analysis such as *p*-hacking.

The third is high researcher degrees of freedom in the search for significance. As Simmons *et al.* (2011) say, ‘undisclosed flexibility in data collection and analysis allows presenting anything as significant.’ This includes flexibility in when to stop collecting data (with peeking at intervals). ‘*P*-hacking’ or ‘data dredging’ is conducting many analyses on the same dataset and just reporting those that were statistically significant but not disclosing these multiple comparisons. By moderately *p*-hacking two real experiments Simmons *et al.* (2011) demonstrated how easy it is to obtain statistically significant evidence for a transparently false hypothesis: that simply listening to a Beatles song can change a person’s real age (Nelson *et al.* 2018; Simmons *et al.* 2011). (For a demonstration of how easy, try the web app *P*-hacker (<http://shinyapps.org/apps/p-hacker/>)). These procedures almost guarantee finding some comparison(s) ‘significant’ but make any conclusions highly likely to be false-positives (Simmons *et al.* 2011, 2018b).

The fourth horseman, HARKing (hypothesizing after the results are known), is looking back at the data, seeing an exciting pattern, and falsely arguing that it was the *a priori* hypothesis question being tested from the beginning (Kerr 1998). It will feel fine to the experimenter as we have a powerful hindsight bias (Chambers 2017; Munafò *et al.* 2017). It’s been likened to firing an arrow and then drawing a target circle round the arrow after it has landed – you can’t miss! HARKing pretends exploratory research is confirmatory research to test a hypothesis declared in advance. All ‘hypothesis testing’ research is at risk of HARKing or *p*-hacking if the analysis is not specified before the experiment starts.

Nelson *et al.* (2018) now suspect that *p*-hacking explains the paradox of how the overwhelming majority of published findings in psychology are statistically significant, despite the overwhelming majority of studies being underpowered and thus unlikely to obtain results that are statistically significant. They suggest that it is failed analyses, not studies, that go into file-drawers and instead, with *p*-hacking, ‘most failed studies are not *missing*. They are published in our journals, masquerading as successes.’

Techniques including funnel-plots to address selective publication of positive results in past research are discussed by Nelson *et al.* (2018). To evaluate the proportion of true effects and indications of likely *p*-hacking in a given set of studies, a technique called *p*-curve (Simonsohn *et al.* 2014, 2019) plots the distribution of reported *p*-values. A ‘*p*-hacking bump’ just below  $p < 0.05$  may indicate attempts to get just under the ‘significance’ line. However, *p*-curve analysis can give an unjustified ‘all ok’. With heterogeneous sets of published papers, a lack of ‘*p*-hacking bump’ and a right-skewed *p*-curve ‘clean bill of health’ is not conclusive evidence that there is no *p*-hacking or that the studies have evidential value (see e.g. (Bishop & Thompson 2016)).

The question about *p*-curve analysis illustrates a problem with meta-analysis in a field like psychology where different teams study different effects and studies are very different, even on a single concept (Nelson *et al.* 2018). The meta-analyst cannot reasonably assess if the original results in each paper were based on errors in data collection, design, or undisclosed flexibility in analysis. The end result of a meta-analysis is only as strong as the weakest studies, and meta-analysis can have its own biases not limited to which studies to include, both factors giving the risk that meta-

analysis exacerbates rather than solves the problems (Ioannidis 2010; Lakens *et al.* 2016; Nelson *et al.* 2018).

The lack of direct replication of experiments in psychology, and across the life-sciences in general, is a key underlying cultural problem. A direct replication seeks to test the repeatability of a previous finding by duplicating the methodology as exactly as possible (Chambers 2017). Whereas physics researchers expect direct replication before accepting new ideas, in psychology out of every 1,000 papers only two are a direct replication and only one of these will be by a different lab (Chambers 2017, p50). Without direct replications, psychology has lost the ability to self-correct. Instead of direct replications, psychology replicates concepts with novel experiments that test a related (but different) idea using a different method (Chambers 2017, pp 13-16, 48-55); for example, testing social priming of stereotypes in different contexts and situations. Arguably, psychology's 'conceptual replications' are not real replications. The original study, which may be erroneous, is not re-tested. Like the rest of psychology, 'conceptual replications' may be the norm in human chemical communication research. The many experiments using androstadienone and/or estratetraenol fit this pattern as experiments test supposed effects without going back to see if they really were pheromones in the first place.

### **(b) Case history: The rise and fall of 'power posing'**

If the 'putative human pheromone' molecules androstadienone and estratetraenol are highly unlikely to have any effect, based on the lack of *any* primary evidence that these are pheromones (Section §3), how is it that 60 papers using androstadienone and/or estratetraenol report positive results?

Could these be 'false positives' and the biased reporting only of positive results?

There are precedents in mainstream psychology for a widely believed phenomenon to be shown to probably be based on false positives, including ones claiming physiological effects. One example is 'power posing', apparently supported by more than 50 papers showing positive results that has ultimately been found to be baseless (Section §7a) (Jonas *et al.* 2017). Power posing is an interesting and attractive idea which has been publicised in a TED talk viewed more than 50 million times: if our confidence is displayed by our physical posture, could our posture affect our behaviour and physiology? Would posing our bodies into a 'confident' 'powerful' pose become 'embodied' and change the way we behave? In the original study by Carney *et al.* (Carney *et al.* 2010), 42 participants were randomly assigned to be posed by the experimenter for two minutes in 'open' expansive 'high-power poses', limbs widespread, or instead 'closed', contractive 'low-power' poses. Participants given 'high-power' poses apparently had raised saliva testosterone levels and lowered stress hormone cortisol levels, and were more willing to take financial risks in a gambling task (Carney *et al.* 2010). Participants also self-reported that they felt more powerful (presented as a check for the manipulation) but the key claim of the paper was that 'embodiment extends beyond mere thinking and feeling, to physiology and subsequent behavioral choices' (Carney *et al.* 2010).

The first problems came with a much higher powered conceptual replication study, with 200 participants, which failed to show any significant effects of power posing on hormonal levels or risk taking behaviour (Ranehill *et al.* 2015). Self-reported 'felt feelings' of power did replicate. Carney *et al.* (2015) responded in the same issue of the journal with a tabulation of 33 apparently successful studies, including Carney *et al.* (2010), which they concluded supported the hypothesis. Simmons and Simonsohn (2017) then did a *p*-curve analysis of the 33 studies, which found an average effect size of zero, suggesting that selective reporting was likely to be the reason for the uniformly positive results in the literature.

Cuddy *et al.* (2018) later responded with a *p*-curve analysis of their own including additional studies (published up to December 2016), suggesting that these supported self-reported ‘feelings of power’ (not the original claims by Carney *et al.* (2010) of changes to physiological measures of hormone levels and risk taking behaviour). However, this *p*-curve analysis itself is likely to be unreliable as its conclusions rely on four outliers with unlikely and extreme low probabilities (Simmons *et al.* 2018a).

Courageously, Dana Carney, first author on the 2010 paper, has now stated that she does not believe ‘power pose’ effects are real, listing problems with the original paper, including *p*-hacking thought acceptable at the time (Carney 2016). Even more important, Carney has helped other researchers conduct a rigorous pre-registered set of replications (see Section §7) that conclude that power posing has no effects on behaviour and physiology, the original 2010 claims, even if it does give a feeling of ‘felt power’.

My reason for covering this story at length is that it is a good example of how a positive literature can uniformly report a non-existent effect. The power posing story shows it is plausible that a positive literature for the ‘putative human pheromone’ molecules could build up without any underlying real phenomenon. It also demonstrates the complexity of teasing out what has been shown and what has not.

## 6. Doing it better: Psychology’s ‘credibility revolution’ and ‘renaissance’

The response of psychology over the last 7 or so years has been swift, innovative, and persuasive. Vazire (2018) argues that rather than describing an ongoing ‘reproducibility crisis’, with the implication that there are no solutions, she suggests instead a ‘credibility revolution’, emphasising the increased reliability and solidity of science that can result from the new approaches outlined below. Nelson *et al.* (2018) similarly argue for ‘psychology’s renaissance’. The science will be better for it – even though it might seem to take longer, with fewer, bigger studies instead of quicker ‘exciting’, but ultimately unreliable, small studies.

A key proposal to reduce bias in data collection and analysis is a move to pre-registration of studies and Registered Reports, described in the next section. In a brief but wide ranging manifesto for reproducible science, Munafò *et al.* (2017) also argue that institutional, funder and other stakeholders need to change incentives for researchers so better behaviour is rewarded: success should come from producing rigorous, transparent, reproducible science not the opposite. While there are costs to improving reproducibility, Poldrack (2019) outlines many benefits, including ones for Early Career Researchers (ECRs). Open science, with sharing of data and software code, is recommended as one solution to small underpowered studies by lone researchers. The Open Science Framework (OSF)([www.osf.io](http://www.osf.io)) is among the platforms offering many tools to facilitate these more collaborative approaches.

## 7. Pre-registration of studies and Registered Reports

A pre-registration is a time-stamped plan for an experiment including data collection and detailed plans for analysis before any data is collected and analysed (Nelson *et al.* 2018; Nosek *et al.* 2018). Pre-registration thwarts *p*-hacking, HARKing, and publication bias. Psychologists currently have two main options for pre-registration: AsPredicted (<http://AsPredicted.org>) and the OSF (<https://osf.io/prereg/>)(Nelson *et al.* 2018). The Registered Report takes the idea a stage further by adding pre-study peer-review to the pre-registration, with an incentive to the researcher of assured publication.

The Registered Report format splits conventional peer-review in half (Chambers 2019a). First, authors submit their plans in a ‘Stage 1’ manuscript which includes an overview of the background literature, preliminary work, theory, hypotheses and proposed methods, including the study procedures and analysis plan (Stage 1 in figure 1) (Chambers 2019a, c; Nosek *et al.* 2018). After incorporation of suggestions from the peer-reviewers (which could include statistical advice received when it matters most, *before* data collection), the journal offers an ‘in-principle acceptance’ (figure 1). Publication decisions at Stage 1 are based on the importance of the research question and quality of the methods proposed to answer that question. With their ‘in-principle acceptance’, the individual researcher pre-registers their Stage 1 manuscript with a recognized repository such as the OSF. They can then collect their data confident in the knowledge that their publication is effectively guaranteed, whether or not the results are ‘significant’, so long as the pre-approved protocol for data collection and analysis is followed (checked by the original reviewers at Stage 2 peer review). Exploratory experiments are still encouraged but are labelled as such and treated as preliminary results which would need a pre-registered experiment to test the idea (Chambers 2019a; Nosek *et al.* 2018). Registered Reports will not be suitable for all research but, for example, there are ways of accommodating sequential experiments, perhaps by presenting the results of such experiments as a Stage 1 manuscript with proposals for experiments to test the resulting predictions (Chambers 2019a). Chambers (2019a) argues that Registered Reports are a plan, not a prison.

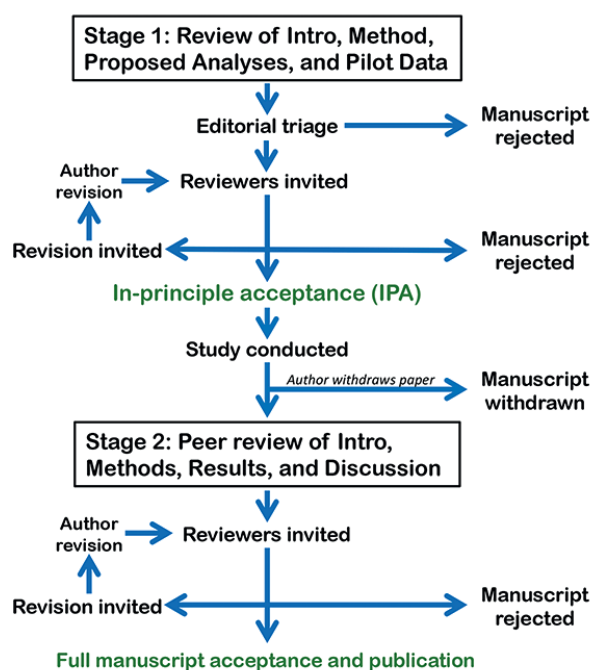


Figure 1. Schematic of the Registered Reports process. Image: Centre for Open Science (<https://cos.io/rr/>) CC BY-ND 4.0

Registered Reports started with one journal, *Cortex*. Six years on, more than 200 science journals, including *Royal Society Open Science* and *Nature Human Behaviour*, have adopted them (Chambers 2019c). Funders and journals are starting to partner to peer review proposals together and then offer

combined funding and In-principle acceptance in the journal to successful proposals (Chambers 2019a). This has the potential to greatly reduce the burden on reviewers as well as improving the design and conduct of experiments.

### **(a) How well are Registered Reports working?**

To date, almost 200 Registered Report articles have been published. The results are encouraging (Chambers 2019a, c). As hoped, it seems positive publication bias is greatly lessened: in one survey, about 60% of Registered Reports, both replications and novel experiments, published negative results (null findings) in marked contrast to the typical 10% in equivalent traditional papers (Allen & Mehler 2019). Researchers will be reassured that Registered Reports are cited, on average, at or above the impact factors of the journals they are published ([tinyurl.com/RR-citations](https://tinyurl.com/RR-citations)) (Chambers 2019c).

But it is early days and there are many things to improve (Chambers 2019a). Hardwicke and Ioannidis (2018) surveyed Registered Reports to February 2018. One surprising problem they identified is a lack of transparency: most agreed protocols were hidden so the final Registered Report could not be compared with the protocol accepted in principle; advocates agree that publication of agreed Stage 1 protocols is essential (Chambers & Mellor 2018) and new public registries will help enable this (Chambers 2019a). Claesen *et al.* (2019) identified a different problem, undeclared deviations from the pre-registration plans, revealed by a study of all articles in *Psychological Science* badged as pre-registration studies in their 2015-2018 time period. These early problems will be sorted out, as they need to be for the system to work as intended.

### **(b) Power posing revisited: resolved by pre-registered replications**

Returning to the question of power posing (Section §5b), in 2017 a special issue of *Comprehensive Results in Social Psychology* was devoted to seven pre-registered studies including direct replications of power posing (all seven studies received advice from Dana Carney, first author of the original 2010 study) (Cesario *et al.* 2017). The seven studies concluded that the original Carney *et al.* (2010) claims of the effects of poses on hormone levels and risk-taking behaviour did not replicate (Jonas *et al.* 2017). The authors commented on the helpful nature of the pre-registration peer-review. For me, it is a powerful example of a difficult controversy resolved by pre-registered studies.

## **8. Taking human chemical communication research forward**

Our sense of smell is undoubtedly an important influence on our behaviour, a view supported by ample evidence in the papers in this special issue [ xx ] and previous reviews by Doty (1981), Schaal and Porter (1991), Shepherd (2004), Wyatt (2014) and McGann (2017). Our olfactory sense is well developed, even if it has often been underestimated. How should we go forward?

As a first step, we should stop using androstadienone and estratetraenol as stimuli since there is no robust evidence that these are ‘human pheromones’, despite their wide use experimentally (see Section §3) (Doty 2014; Wyatt 2015; Wysocki & Preti 2004). There is a real opportunity cost: every experiment using these molecules is scientist-time or funding not going to answer real questions about human chemical communication. I’ll return later to the odours we should investigate instead.

I have argued previously (Wyatt 2015) that we should adopt pre-registration of studies in human chemical communication. We are now in an even better position to learn from progress in sister fields in psychology and other life sciences. The rest of this review offers some positive proposals and methods for researching human pheromones and olfactory cues.

### (a) Adopt Pre-registration and/or Registered Reports

The single change with the most impact would be the adoption of pre-registration and/or Registered Reports for the majority of studies (Chambers 2017; Chambers 2019a; Munafò *et al.* 2017). Perhaps one of the most surprising conclusions from psychology's 'reproducibility crisis' is that our routine good-practice precautions of randomization and running experiments double-blind are essential but are *not enough* to ensure the removal of bias. As explored in Section §5a, gathering the data is just the beginning of potential experimenter choices in analysis that can lead to bias (for a comprehensive catalogue of potential bias, and some ways to reduce them, see [www.catalogofbias.org](http://www.catalogofbias.org), Bishop (Bishop 2020), and Chambers (Chambers 2017)).

Whether we are looking at cues or pheromones, whatever the type of experiments, like other life scientists we are too good at getting 'interesting' results from data. With cues we are looking for subtle effects which make it all the more important to use pre-registration. For functional magnetic resonance imaging (fMRI) too, given the enormous degrees of analytical flexibility with the interpretation of fMRI and other neuroimaging data, pre-registration of the neuroimaging analysis is especially important (Gorgolewski & Poldrack 2016; Poldrack *et al.* 2017).

We could learn from the ways that other areas of science are adopting Registered Reports and other improved practices. For example, cognitive neuroscience researchers recently discussed adoption, in a special forum in the journal *Cortex* with an introduction by Chambers (2019b). Should we convene a similar discussion among researchers in human chemical communication, in a journal such as *Chemical Senses*?

### (b) What do we really know? Re-assessing our existing literature

'Identifying these realities—that researchers engage in *p*-hacking and that *p*-hacking makes it trivially easy to accumulate significant evidence for a false hypothesis—opened psychologists' eyes to the fact that many published findings, and even whole literatures, could be false positive' Nelson *et al.* (2018)p513

Sadly, we cannot assume that the conclusions of papers that we traditionally cite about different aspects of human chemical communication are reliable. We need to be more careful with what we cite in our introductions and discussions. When I have been assessing papers for inclusion in my own literature reviews, though I check for controls and whether the study was done blind, I tend to assume that the peer review system has worked. As mentioned earlier, high impact journals may be less reliable than lower ranked ones (Brembs 2018), and work published there certainly can be fallible as Camerer *et al* (2018) demonstrated.

A problem, right across the life-sciences, is that even when later studies show that a result was probably a false-positive, citations to debunked papers or concepts continue undiminished (for example, only 5% of authors citing retracted biological science papers indicated any awareness that the cited article had been retracted or subject to a matter of concern) (Neale *et al.* 2010)). We currently have no system in biology or behavioural sciences/psychology for systematic links between papers and their replications, failed or successful, and refutations (discussed in e.g. Huber *et al.* 2019). PubPeer.com (Bishop 2018) provides one mechanism for alerts to comments added on individual papers, which could include links to refuting papers, but it is hardly used in chemical communication literature (for example, a search for 'androstadienone' brings up no records, searched 18 October 2019). Perhaps I (we) should be using it more.



Among the tools to assess the reliability of our current knowledge are meta-analysis and replication.

### i. Meta-analysis of existing studies

Some meta-analysis (see Section §5a) has been done on studies of human olfactory cues, for example on the influence of menstrual cycle on odour-based mate choice (see de Groot and Smeets (2017) for references). . In an analysis of the human fear chemosignaling literature, de Groot and Smeets (2017) combined regular meta-analysis, which showed evidence for a small-to-moderate effect size, with p-curve analysis and p-uniform, another tool, which showed evidence diagnostic of a true effect and no evidence of publication bias. However, a meta-analysis is only as reliable as the source studies (shown by many positive meta-analyses of studies of the now disproved links between serotonin transporter gene 5-HTTLPR and depression (Alexander 2019)).

### ii. Replications

Based on the rest of psychology we can perhaps anticipate that 30-40% or more of studies in our field would not replicate. Replicating studies is hard work and can take longer than the original study, especially when the power of the experiments is increased by having many more subjects involved. Which studies should we prioritize for replication? Frequently cited and influential studies perhaps. One way of identifying candidate studies for replication may be crowd-sourced assessment. This is suggested by the second part of the study by Camerer *et al* (2018) (Section §5a) into the reproducibility of social science studies published in *Nature* and *Science*. In parallel to the replication of the previously published experiments, a panel of several hundred volunteers from the scientific community was recruited to evaluate the already published version of the papers. In surveys and a kind of ‘prediction market’, the scientist volunteers, by reading the paper and the plans for replication, were able to predict (with impressive accuracy) how likely a paper would be to replicate (before the replication was carried out). Decisions by a small number of reviewers or a committee risk personal bias. Here the pooled assessment was by hundreds of scientists each making a separate observation independently. Why didn’t the original reviewers and editors detect that some studies were unlikely to replicate, before passing the studies for publication. Was it that they, like all of us, were seduced by the exciting, interesting story that they wanted to believe?

One problem for replications is that many journals will not publish them, successful or unsuccessful, something which impacts early career researchers especially (Allen & Mehler 2019). In a pioneering policy, the Psychology and Cognitive Neuroscience section of *Royal Society Open Science* guarantees to publish close replications of any article published in the journal (the ‘Pottery Barn’ principle, ‘you broke it, you sort it’) and close replications of articles from most other major journals too. Let’s hope that other journals will adopt this policy, especially when the replication is of a study previously published in that journal (‘prestigious’ journals are notoriously reluctant to do this). Funders also need to provide grants which include funding for replications.

A further problem is deciding what makes a successful replication or reproduction. Replications can fail to reproduce the results of the original study for a variety of reasons (Bishop 2020; Bishop 2018; Goodman *et al.* 2016; Nelson *et al.* 2018; Open Science Collaboration 2015). These include known or unknown differences between the replication and original study. It can be hard to know if the original study was a false positive or if the replication is a false negative (though usually the replication will be designed to be statistically more powerful, which helps). We are at an early stage in understanding how best to carry out and interpret reproducibility (Bishop 2018; Goodman *et al.* 2016; Nelson *et al.* 2018).

Some of the complexities of replication are illustrated by studies of human tears as chemical communication. In the original paper published in *Science* in 2011, Gelstein *et al.* (2011) reported that ‘women’s emotional tears contain a chemosignal that reduces sexual arousal in men’ (p. 230). Prompted by this study, Gračanin *et al.* (2017a) tested the effect of women’s emotional tears on men but used attractive female face and body photographs and showed no effects of the tears in three different conceptual replications (using male subjects’ ratings of the photographs and measures of pro-social behaviour). The senior author of the 2011 paper, Sobel countered (Sobel 2017) with detailed objections to which Gračanin *et al.* (2017b) replied. My observation is that Gelstein *et al.* (2011) claimed a strong effect and implied this worked ‘in the wild’; they did not claim this was a fragile effect only present in the lab when looking at particular emotionally neutral faces or watching sad, happy or neutral films. Gračanin *et al.* (2017a)p11 concluded ‘if there is any substance in female’s tears that has a dampening effect on the sexual arousal of males, this influence is very modest at best and certainly does not always impact every male in his sexual functioning.’

### (c) The challenges of working with human chemical communication

Studies on human chemical communication share all the challenges facing psychologists studying any aspect of human behaviour. However, in addition, dealing with odours is more complex than visual or sound stimuli – there is no format like MP3 for recording or replaying a smell. Researchers have built up a body of good practice for isolating, storing, and delivering olfactory stimuli (see other papers in this issue and Doty 2015; Drea *et al.* 2013; Drea 2015; Frumin & Sobel 2013; Keller & Vosshall 2004; Kjeldmand *et al.* 2011; Lapid & Hummel 2013; Lundström *et al.* 2010; Miyazawa *et al.* 2009; Oleszkiewicz *et al.* 2018). The extremely low active concentrations present further challenges for both measurement and then delivery. It also needs to be kept in mind that our understanding of olfaction as a sense has lagged behind the senses of vision and hearing so the unknowns may be greater (Barwich 2020; Barwich 2016; McGann 2017). Olfaction shares with the other senses, perhaps more so, the added complication of differences in human culture mediated through language and experience (Arshamian *et al.* this issue).

Whether the search is for cues or pheromones, it needs to start with a behaviour or situation for which there is some evidence that smell may be involved. This can form the basis of a bioassay (Section §2b) when later exploring which molecules might be involved. With human behaviour, subtle responses are likely (though this is also a problem for studies of mice and other mammals too). A nice example is Frumin *et al.* (2015)’s study of unconscious sniffing of hands after handshaking. Subjects unaware of the purpose of the study were covertly videoed after staged but ‘natural’ handshakes with a gloved or ungloved experimenter. Video analysis focussed on the proximity of the subject’s hands near their nose in the 60 seconds before and after the handshake. The only improvement might have been pre-registration of the experiments and analysis (but that goes for the majority of studies cited in this review).

Developing a reliable and repeatable quantifiable measure (bioassay), such as the hand-nose proximity in the handshaking study above, biologically relevant to the response being investigated is a key task for experimenters (Brown & Bolivar 2018; Wyatt 2014, Chapter 2). The suckling response of human babies to mammary secretions of mothers is another nice example (Doucet *et al.* 2009, 2012; Schaal *et al.* this issue). Cultural and ‘ecological relevance’ is important for human measures. Saxton *et al.* (2008)’s study of speed-dating was a good idea (even if they were, to my mind, using the wrong molecule, androstadienone). Ultimately, we do need the effects to work ‘in the wild’.



We would anticipate that cues would be context dependent, but pheromones are too. Pheromones, by definition, elicit stereotyped behavioural and/or physiological responses in the receiver, in all animals including mammals, but these responses are also modulated by context, time of day, and many other factors including the receiver's genetics, age, sex, hormonal state, dominance status, and recent experience (Wyatt 2014, p206-9). For studies of cues or pheromones, the experimenter's dilemma is how to control for these variables (hence the detailed protocols) without completely reducing the generalizability of the results, if the effect then only works in the lab in very limited conditions.

#### **(d) What molecules and from where**

I describe practical details of how to approach human chemical communication in Wyatt (Wyatt 2015). The first task is to identify well characterized phenomena influenced by olfactory stimuli. Initially we should use the natural odours/secretions under investigation (e.g. possible fear-associated odours (de Groot this issue)). Such studies have the advantage that, even though the molecules are unknown, the concentrations used are natural. The next task will be to identify the molecule(s) involved. Cues may be highly variable between individuals (for example to do with recognition of siblings) or cues could be the same in every individual, such as acetone in the breath of anyone in diabetic ketoacidosis (Shirasu & Touhara 2011).

Pheromones occur superimposed on the background of hundreds of molecules making up a highly variable individual odour profile, from many sources, including those secreted by the organism, from bacteria, picked up from conspecifics, and from diet (Natsch & Emter this issue; Wyatt 2010; 2014, pp 2-16, 284-291). Subtractive techniques to reveal candidate molecules against this background could use comparisons of odour profiles between prepubescent children, women and men (Wyatt 2015). We still lack robust studies of odour emission at different human life stages and only a few large scale studies look at adult male and females (e.g. Penn *et al.* 2007). Any future proposed pheromones will need to satisfy the criteria outlined in Section §2b and Box 1.

New chemical techniques offer ways of analysing the volatile odours given off by humans in different contexts in real time (Roberts this issue; Williams *et al.* 2016). This will potentially free researchers from the limitations of having to take samples from particular parts of the body over a period of time, typically collecting odours on cotton pads under the armpits for example, which averages the collection over many hours.

On the one hand, the Darwinian approach of treating humans as just another primate/mammal is fruitful. However, caution is needed when using other species' pheromones as evidence for the likelihood that pheromones of a particular kind exist in a second species or that a particular molecule is likely to be a pheromone in a second species. That mice use a male tear-gland protein, ESP1, as a sex pheromone is no indication that tears will be important in human chemical communication though this was the justification for (Gelstein *et al.* 2011)p226. Similarly, 'copulins', not even established as sex pheromones in rhesus monkeys (Wyatt 2015), are still being studied as 'human sex pheromones' (e.g. Williams & Apicella 2017)). Conversely, a particular molecule being a pheromone in one species does not rule it out being a pheromone (or a component of one) in another species but being a pheromone in one species does not make it more likely that the same molecule is a pheromone in another species. For example, androstenone and androstenol are thought to be the male pig sex pheromone (though see (Doty 2010, 2014)) but the presence in small quantities in human armpits is no evidence that these are human pheromones (Doty 2010, 2014).

## 9. Conclusions

I ended my 2015 review (Wyatt 2015) by writing ‘It may be that we will find that there are no pheromones in humans. But we can be sure that we shall never find anything if we follow the current path. We need to start again.’ I still have the same view, cautiously optimistic. I am even more persuaded that humans respond to olfactory cues, but I can see that it will be challenging to demonstrate these beyond doubt. For both pheromones and olfactory cues, it will be a separate and difficult task to identify the molecules involved. As we go forward, I am full of admiration for the scientists who take on these challenges.

## Acknowledgements

Thanks to Jessie Baldwin, Ann-Sophie Barwich, Heather Eisthen, Jasper de Groot, Jan Havlíček, Jonas Olofsson, Andreas Natsch, Benoist Schaal, John Williams, and two anonymous referees for their helpful comments. Any remaining errors are mine of course.

## Prior Versions

An earlier version of this article was posted as a preprint at <https://peerj.com/manuscripts/40453/>.

## Supplemental File

Primary research papers using androstadienone and/or estratetraenol as the stimulus to human subjects 1945-2019.

## References

- Alexander, S. 2019. 5-HTTLPR: A pointed review. *Slate Star Codex* [Online]. Available from: <https://web.archive.org/web/20200105195137/https://slatestarcodex.com/2019/05/07/5-http-a-pointed-review/> [Accessed 05 January 2020].
- Allen, C & Mehler, DMA (2019) Open science challenges, benefits and tips in early career and beyond. *PLoS Biology* 17: e3000246.
- Arshamian, A, Manko, P & Majid, A (this issue) Why communication and simulation of olfaction is different from other sensory stimuli. *Phil Trans Roy Soc B*.
- Bargh, JA, Chen, M & Burrows, L (1996) Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology* 71: 230.
- Barwich, A-S (2020) *Smellosophy - What the nose tells the mind*. Cambridge, Mass: Harvard University Press.
- Barwich, AS (2016) What is so special about smell? Olfaction as a model system in neurobiology. *Postgrad Med J* 92: 27-33.
- Begley, CG & Ellis, LM (2012) Raise standards for preclinical cancer research. *Nature* 483: 531.
- Bishop, DV & Thompson, PA (2016) Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ* 4: e1715.
- Bishop, DV (2019) Rein in the four horsemen of irreproducibility. *Nature* 568: 435-435.
- Bishop, DV (2020) The psychology of experimental psychologists: Overcoming cognitive constraints to improve research: The 47th Sir Frederic Bartlett Lecture. *Q J Exp Psychol (Hove)* 73: 1-19.
- Bishop, DVM (2018) Fallibility in Science: Responding to Errors in the Work of Oneself and Others. *Advances in Methods and Practices in Psychological Science* 1: 432-438.
- Brembs, B (2018) Prestigious science journals struggle to reach even average reliability. *Front Hum Neurosci* 12: 37.

- 735 Brown, RE & Bolivar, S (2018) The importance of behavioural bioassays in neuroscience. *Journal of Neuroscience Methods* 300: 68-76.
- Button, KS, *et al.* (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews in Neuroscience* 14: 365-376.
- Camerer, CF, *et al.* (2018) Evaluating the replicability of social science experiments in *Nature* and  
740 *Science* between 2010 and 2015. *Nature Human Behaviour* 2: 637-644.
- Cardé, RT (2015) Multi-cue integration: How female mosquitoes locate a human host. *Current Biology* 25: R793-795.
- Carney, DR, Cuddy, AJ & Yap, AJ (2010) Power posing: brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychol Sci* 21: 1363-1368.
- 745 Carney, DR, Cuddy, AJ & Yap, AJ (2015) Review and summary of research on the embodied effects of expansive (vs. contractive) nonverbal displays. *Psychol Sci* 26: 657-663.
- Carney, DR 2016 *My position on "Power Poses"* [Online]. Available:  
[https://web.archive.org/web/20190723070448/http://faculty.haas.berkeley.edu/dana\\_carney/pdf\\_My%20position%20on%20power%20poses.pdf](https://web.archive.org/web/20190723070448/http://faculty.haas.berkeley.edu/dana_carney/pdf_My%20position%20on%20power%20poses.pdf) [Accessed 23 July 2019].
- 750 Cesario, J, Jonas, KJ & Carney, DR (2017) CRSP special issue on power poses: what was the point and what did we learn? *Compr Results Soc Psychol* 2: 1-5.
- Chambers, C (2017) *The seven deadly sins of psychology : a manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.
- Chambers, CD & Mellor, DT (2018) Protocol transparency is vital for registered reports. *Nature*  
755 *Human Behaviour* 2: 791-792.
- Chambers, CD (2019a) What's next for Registered Reports? *Nature* 573: 187-189.
- Chambers, CD (2019b) The battle for reproducibility over storytelling. *Cortex* 113: A1-A2.
- Chambers, CD (2019c) The registered reports revolution Lessons in cultural reform. *Significance* 16: 23-27.
- 760 Claesen, A, Gomes, SLBT & Tuerlinckx, F (2019) Preregistration: Comparing dream to reality. *PsyArXiv*.
- Cuddy, AJC, Schultz, SJ & Fosse, NE (2018) P-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effects: Reply to Simmons and Simonsohn (2017). *Psychol Sci* 29: 656-666.
- 765 Curran, T (2018) Reproducibility of academic preclinical translational research: lessons from the development of Hedgehog pathway inhibitors to treat cancer. *Open Biol* 8.
- de Groot, JH, *et al.* (2012) Chemosignals communicate human emotions. *Psychological science* 23: 1417-1424.
- de Groot, JH (this issue) Three degrees of fear: the chemical communication of emotion intensity.  
770 *Phil Trans Roy Soc B*.
- de Groot, JHB & Smeets, MAM (2017) Human fear chemosignaling: Evidence from a meta-analysis. *Chemical Senses* 42: 663-673.
- Doty, RL (1981) Olfactory communication in humans. *Chemical Senses* 6: 351-376.
- Doty, RL (2010) *The great pheromone myth*. Baltimore, MD: Johns Hopkins University Press.
- 775 Doty, RL (2014) Human pheromones: Do they exist? In: Mucignat-Caretta, C (ed.) *Neurobiology of chemical communication*. pp. 535-560. Boca Raton, FL: CRC Press.
- Doty, RL (ed.) (2015) *Handbook of olfaction and gustation. 3rd Edition*. Chichester: Wiley-Blackwell.
- Doucet, S, *et al.* (2009) The secretion of areolar (Montgomery's) glands from lactating women elicits selective, unconditional responses in neonates. *PLoS ONE* 4: e7579.
- 780 Doucet, S, *et al.* (2012) An overlooked aspect of the human breast: areolar glands in relation with breastfeeding pattern, neonatal weight gain, and the dynamics of lactation. *Early Human Development* 88: 119-128.
- Doyen, S, *et al.* (2012) Behavioral priming: It's all in the mind but whose mind? *PLoS ONE* 7.

- 785 Drea, CM, *et al.* (2013) The "secret" in secretions: methodological considerations in deciphering  
primate olfactory communication. *American Journal of Primatology* 75: 621-642.
- Drea, CM (2015) D'scent of man: A comparative survey of primate chemosignaling in relation to  
sex. *Hormones and Behavior* 68: 117-133.
- 790 Fanelli, D (2010) "Positive" results increase down the hierarchy of the sciences. *PLoS ONE* 5:  
e10068.
- Ferdenzi, C, *et al.* (2016) Androstadienone's influence on the perception of facial and vocal  
attractiveness is not sex specific. *Psychoneuroendocrinology* 66: 166-175.
- Ferkin, MH, *et al.* (1994) Attractiveness of male odors to females varies directly with plasma  
testosterone concentration in meadow voles. *Physiology & Behavior* 55: 347-353.
- 795 Frumin, I & Sobel, N (2013) An assay for human chemosignals. In: Touhara, K (ed.) *Pheromone  
Signaling*. pp. 373-394. New York: Humana Press.
- Frumin, I, *et al.* (2015) A social chemosignaling function for human handshaking. *Elife* 4.
- Gelstein, S, *et al.* (2011) Human tears contain a chemosignal. *Science* 331: 226-230.
- Goodman, SN, Fanelli, D & Ioannidis, JPA (2016) What does research reproducibility mean?  
800 *Science Translational Medicine* 8: 341ps312-341ps312.
- Gorgolewski, KJ & Poldrack, RA (2016) A practical guide for improving transparency and  
reproducibility in neuroimaging research. *PLoS Biology* 14: e1002506.
- Gračanin, A, *et al.* (2017a) Chemosignalling effects of human tears revisited: Does exposure to  
female tears decrease males' perception of female sexual attractiveness? *Cognition and  
805 Emotion* 31: 139-150.
- Gračanin, A, Vingerhoets, AJ & van Assen, MA (2017b) Response to comment on  
"Chemosignalling effects of human tears revisited: Does exposure to female tears decrease  
males' perception of female sexual attractiveness?". *Cogn Emot* 31: 158-159.
- Hardwicke, TE & Ioannidis, JPA (2018) Mapping the universe of registered reports. *Nature Human  
810 Behaviour* 2: 793-796.
- Hare, RM, *et al.* (2017) Putative sex-specific human pheromones do not affect gender perception,  
attractiveness ratings or unfaithfulness judgements of opposite sex faces. *Royal Society Open  
Science* 4: 160831.
- Haselton, MG & Gildersleeve, K (2016) Human ovulation cues. *Current Opinion in Psychology* 7:  
815 120-125.
- Havlíček, J, *et al.* (2006) Non advertized does not mean concealed: body odour changes across the  
human menstrual cycle. *Ethology* 112: 81-90.
- Havlíček, J, *et al.* (2010) Current issues in the study of androstenes in human chemosignaling. In:  
Gerald, L (ed.) *Pheromones*. pp. 47-81. London: Academic Press.
- 820 Havlíček, J, Winternitz, J & Roberts, SC (this issue) MHC-associated odour preferences and human  
mate choice: near and far horizons. *Phil Trans Roy Soc B*.
- Huber, DE, Potter, KW & Huszar, LD (2019) Less "story" and more "reliability" in cognitive  
neuroscience. *Cortex* 113: 347-349.
- Hummer, TA & McClintock, MK (2009) Putative human pheromone androstadienone attunes the  
mind specifically to emotional information. *Hormones and Behavior* 55: 548-559.
- 825 Huoviala, P & Rantala, MJ (2013) A putative human pheromone, androstadienone, increases  
cooperation between men. *PLoS ONE* 8: e62499.
- Ioannidis, JP (2010) Meta-research: The art of getting it wrong. *Res Synth Methods* 1: 169-184.
- Ishii, KK & Touhara, K (2018) Neural circuits regulating sexual behaviors via the olfactory system  
830 in mice. *Neuroscience Research*.
- Jacob, S & McClintock, MK (2000) Psychological state and mood effects of steroidal chemosignals  
in women and men. *Hormones and Behavior* 37: 57-78.
- Jonas, KJ, *et al.* (2017) Power poses – where do we stand? *Comprehensive Results in Social  
Psychology* 2: 139-141.

- 835 Karlson, P & Lüscher, M (1959) 'Pheromones': a new term for a class of biologically active substances. *Nature* 183: 55-56.
- Keller, A & Vosshall, LB (2004) Human olfactory psychophysics. *Current Biology* 14: R875-R878.
- Kerr, NL (1998) HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review* 2: 196-217.
- 840 Kjeldmand, L, Salazar, LTH & Laska, M (2011) Olfactory sensitivity for sperm-attractant aromatic aldehydes: a comparative study in human subjects and spider monkeys. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* 197: 15-23.
- Klein, RA, *et al.* (2018) Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science* 1: 443-490.
- 845 Lakens, D, Hilgard, J & Staaks, J (2016) On the reproducibility of meta-analyses: six practical recommendations. *BMC Psychol* 4: 24.
- Lapid, H & Hummel, T (2013) Recording odor-evoked response potentials at the human olfactory epithelium. *Chemical Senses* 38: 3-17.
- 850 Liberles, SD (2014) Mammalian pheromones. *Annual Review of Physiology* 76: 151-175.
- Lobmaier, JS, *et al.* (2018) Accumulating evidence suggests that men do not find body odours of human leucocyte antigen-dissimilar women more attractive. *Proc Biol Sci* 285.
- Lundström, JN, *et al.* (2010) Methods for building an inexpensive computer-controlled olfactometer for temporally-precise experiments. *International Journal of Psychophysiology* 78: 179-189.
- 855 Maynard Smith, J & Harper, D (2003) *Animal signals*. Oxford: Oxford University Press.
- McClintock, MK (1971) Menstrual synchrony and suppression. *Nature* 229: 244-245.
- McGann, JP (2017) Poor human olfaction is a 19th-century myth. *Science* 356.
- Miyazawa, T, *et al.* (2009) Methodological factors in odor detection by humans. *Chemosensory Perception* 2: 195-202.
- 860 Monti-Bloch, L & Grosser, BI (1991) Effect of putative pheromones on the electrical activity of the human vomeronasal organ and olfactory epithelium. *Journal of Steroid Biochemistry and Molecular Biology* 39: 573-582.
- Munafò, MR, *et al.* (2017) A manifesto for reproducible science. *Nature Human Behaviour* 1: 0021.
- Natsch, A & Emter, R (this issue) The specific biochemistry of human axilla odour formation viewed in an evolutionary context. *Phil Trans Roy Soc B*.
- 865 Neale, A, Dailey, R & Abrams, J (2010) Analysis of citations to biomedical articles affected by scientific misconduct. *Science and Engineering Ethics* 16: 251-261.
- Nelson, LD, Simmons, J & Simonsohn, U (2018) Psychology's renaissance. *Annual Review of Psychology* 69: 511-534.
- 870 Nosek, BA, *et al.* (2018) The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America* 115: 2600-2606.
- Nuzzo, R (2015) Fooling ourselves. *Nature* 526: 182-185.
- Oleszkiewicz, A, *et al.* (2018) The confounding effect of background odors on olfactory sensitivity testing. *Journal of Neuroscience Methods* 306: 88-91.
- 875 Olsson, MJ (this issue) Cues of sickness in body odour. *Phil Trans Roy Soc B*.
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349.
- Penn, DJ, *et al.* (2007) Individual and gender fingerprints in human body odour. *Journal of the Royal Society Interface* 4: 331-340.
- 880 Plesser, HE (2018) Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics* 11.
- Poldrack, RA, *et al.* (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience* 18: 115.
- Poldrack, RA (2019) The costs of reproducibility. *Neuron* 101: 11-14.

- 885 Ranehill, E, *et al.* (2015) Assessing the robustness of power posing: no effect on hormones and risk tolerance in a large sample of men and women. *Psychol Sci* 26: 653-656.
- Roberts, SC (this issue) Tracking context-dependent odour changes in real time. *Phil Trans Roy Soc B*.
- 890 Saxton, TK, Little, AC & Roberts, SC (2008) Ecological validity in the study of human pheromones. In: Hurst, JL, Beynon, RJ, Roberts, SC & Wyatt, TD (eds.) *Chemical Signals in Vertebrates 11*. pp. 111-120. Springer New York.
- Schaal, B & Porter, RH (1991) Microsmatic humans revisited - the generation and perception of chemical signals. *Advances in the Study of Behavior* 20: 135-199.
- Schaal, B, *et al.* (this issue) Olfaction in parent-to-offspring communication and beyond: a critical review and future directions *Phil Trans Roy Soc B*.
- 895 Shepherd, GM (2004) The human sense of smell: are we better than we think? *PLoS Biology* 2: 572-575.
- Shirasu, M & Touhara, K (2011) The scent of disease: volatile organic compounds of the human body related to disease and disorder. *Journal of Biochemistry* 150: 257-266.
- 900 Simmons, J, Nelson, L & Simonsohn, U. 2018a. [66] Outliers: Evaluating a new p-curve of power poses. *DataColada.org* [Online]. Available from: <http://datacolada.org/66> [Accessed 03 August 2019].
- Simmons, JP, Nelson, LD & Simonsohn, U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22: 1359-1366.
- 905 Simmons, JP & Simonsohn, U (2017) Power posing: P-curving the evidence. *Psychol Sci* 28: 687-693.
- Simmons, JP, Nelson, LD & Simonsohn, U (2018b) False-positive citations. *Perspect Psychol Sci* 13: 255-259.
- 910 Simonsohn, U, Nelson, LD & Simmons, JP (2014) P-curve: a key to the file-drawer. *J Exp Psychol Gen* 143.
- Simonsohn, U, Nelson, LD & Simmons, JP (2019) P-curve won't do your laundry, but it will distinguish replicable from non-replicable findings in observational research: Comment on Bruns & Ioannidis (2016). *PLoS ONE* 14: e0213454.
- 915 Smith, TD, Laitman, JT & Bhatnagar, KP (2014) The shrinking anthropoid nose, the human vomeronasal organ, and the language of anatomical reduction. *The Anatomical Record* 297: 2196-2204.
- Sobel, N (2017) Revisiting the revisit: added evidence for a social chemosignal in human emotional tears. *Cogn Emot* 31: 151-157.
- 920 Stern, K & McClintock, MK (1998) Regulation of ovulation by human pheromones. *Nature* 392: 177-179.
- Tirindelli, R, *et al.* (2009) From pheromones to behavior. *Physiological Reviews* 89: 921-956.
- Vazire, S (2018) Implications of the credibility revolution for productivity, creativity, and progress. *Perspect Psychol Sci* 13: 411-417.
- 925 Wang, D, *et al.* (2018) Irreproducible text-book “knowledge”: The effects of color bands on zebra finch fitness. *Evolution* 72: 961-976.
- Williams, J, *et al.* (2016) Cinema audiences reproducibly vary the chemical composition of air during films, by broadcasting scene specific emissions on breath. *Sci Rep* 6: 25464.
- Williams, MN & Apicella, C (2017) Synthetic copulin does not affect men’s sexual behavior. *Adaptive Human Behavior and Physiology* 4: 121-137.
- 930 Wyatt, TD (2010) Pheromones and signature mixtures: defining species-wide signals and variable cues for identity in both invertebrates and vertebrates. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* 196: 685-700.

- 935 Wyatt, TD (2014) *Pheromones and animal behavior: Chemical signals and signatures*, 2 edn.  
Cambridge: Cambridge University Press.
- Wyatt, TD (2015) The search for human pheromones: the lost decades and the necessity of returning to first principles. *Proceedings of the Royal Society of London B: Biological Sciences* 282: 2014.2994.
- Wyatt, TD (2017) Pheromones. *Current Biology* 27: R739–R743.
- 940 Wysocki, CJ & Preti, G (2004) Facts, fallacies, fears, and frustrations with human pheromones. *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology* 281A: 1201-1211.
- Yong, E (2012) Replication studies: Bad copy. *Nature* 485: 298-300.
- Yong, E. 2018 Psychology's replication crisis is running out of excuses. *The Atlantic* [Online].  
945 Available: <https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/> [Accessed 15 August 2019].