Be sceptical when researchers claim sex differences

Judy Robertson

It's hard to avoid news stories about differences between men and women. As I write this, for example, there is a hot debate about an ill-judged memo written by a Google employee in which he claimed that women are under-represented in technology jobs because men and women have different traits. In his view, women have more of an interest in people and aesthetics, while men tend to be attracted to coding and systematizing.

Today there are also various media reports of a neuroimaging study which claims that women are better at empathising because they have increased prefrontal cortex blood flow in comparison to men, and the Daily Mail tells me that viruses target men because they (the viruses, not the Daily Mail) see men as weaker.

In the last two decades, the number of journal articles about sex differences has doubled, and the number of articles in news media has increased five-fold (Maney, 2016). In 2013, the National Institutes of Health in the US introduced a policy which mandated the inclusion of both sexes in preclinical research with animals, tissues and cells, as well as a requirement to disaggregate the data by sex and compare the sexes where possible (Clayton & Collins, 2014). This will no doubt have increased the number of headlines about the inability of lady mice to read maps, and the sexual prowess of male adrenal gland cells.

If you join the dots between an increase in studies about sex differences and the criticisms about the major flaws in commonly used statistical methods across the sciences and social sciences (Ziliak & McCloskey, 2008; Wasserstein & Lazar, 2016), it gets rather worrying. Ioannidis's classic paper revealed that most published research findings are false (Ioannidis, 2005). There is a set of prevalent but poor research design and analytical practices which contribute to the publication of misleading results – Simmons et al. have demonstrated that if you're over-flexible with your data collection you can get significant results for just about anything, including support for the hypothesis that listening to the song "When I'm 64" literally does make you younger (Simmons et al., 2011).

Another problem is that once a study with misleading claims is published, it's not very likely to be checked through replication. The Open Science Collaboration rocked psychology by repeating 100 landmark studies and finding that only 36% of the replications confirmed statistically significant results (Open Science Collaboration, 2015). If you prefer a Bayesian perspective on this project, see Etz & Vandekerckhove (2016). It is also known that lab-based gender studies have particularly low validity, mostly because of their small effect sizes (Mitchell, 2012). It's not just psychologists who have these problems. For example, in a review of published studies of claims of sex differences for genetic effects, Ioannidis and colleagues discovered that it was uncommon for the studies to document good internal and external validity. Of over 432 claims of sex differences, only 60 had internal validity, and only one of these claims had been consistently replicated in two other studies.

So, we know that the scientific literature is riddled with false results due to statistical

...of over 432 claims of sex differences, only 60 had internal validity, and only one of these claims had been consistently replicated in two other studies...

flaws and publication bias. That's enough of a reason in itself to be sceptical of sex differences reported in journal articles, never mind news articles which have been garbled by journalists who don't have specific training in science reporting. In addition to this general background error rate, there has been a historical bias towards looking for scientific results which confirm stereotypical beliefs about men and women, as described by Angela Saini in her book *Inferior* (Saini, 2017). Cordelia Fine's books *Delusions of Gender* (Fine, 2010) and *Testosterone*

(Fine, 2017) are witty and coruscating well-argued explanations of the biases and flawed reasoning which lead to "men are from Mars, women are from Venus" type arguments and dubious "Just so!" stories about how our modern day behaviour can be explained (or excused) through deep biological urges shaped by evolutionary pressures. (Fine, 2010, 2017).

Such arguments often share a flawed line of reasoning, as elucidated by Donna Maney:

Assertions are based on the following logic: (i) a structure (or hormone) we'll call 'X' differs between men and women; (ii) X is related to a behaviour we'll call 'Y'; (iii) men and women differ in Y; therefore, the sex difference in X causes the sex difference in Y. This argument is invalid because it invokes the false cause fallacy – a sex difference in Y cannot be deduced to depend on X. In addition to being invalid, the argument is also often unsound in that rarely are all three premises supported. (Maney, 2016, p. 3)

In a recent article in *Frontiers in Human Neuroscience*, Rippon and colleagues point out that scientists often have a layperson's understanding of gender scholarship, writing:

Sex/gender NI [neuroimaging] research currently often appears to proceed as if a simple essentialist view of the sexes were correct: that is, as if sexes clustered distinctively and consistently at opposite ends of a single gender continuum, due to distinctive female vs. male brain circuitry, largely fixed by a sexually-differentiated genetic blueprint. (Rippon et al., 2014, p. 1)

There is not such a simple dichotomy. As Fine argues, there are not natural “essences” of men and women, naturally occurring characteristics which are determined by biological factors and invariant to history and culture.

The genetic and hormonal components of sex certainly influence brain development and function ... – sex is just one of many interacting factors. We are an adapted species of course, but also unusually adaptable. Beyond the genitals, sex is surprisingly dynamic, and not just open to influence from gender constructions, but reliant on them. Nor does sex inscribe us with male brains or female brains, or with male natures and female natures. There are no essential male or female characteristics. (Fine, 2017)

Indeed, meta-analyses indicate that considerable support for the gender similarity hypothesis that males and females are similar on most, but not all, psychological variables (Hyde, 2005; Hyde, 2014): in a review of 46 meta-analyses, Hyde found that 78% of the gender differences reported in previous studies were small or very close to zero (Hyde, 2005). Men and women are more alike than stereotypes – and news reports – would have us believe.

It could be that in focusing the analysis on the binary of whether a difference exists or not, we are falling prey to what Maney terms the “methodological fallacy”, the belief that “with respect to any trait the sexes are either fundamentally different or they are the same” (Maney, 2016, p.

2). Perhaps “Is there a difference?” is a misleading question, and both “yes” and “no” are wrong answers. Better questions are: “How much do they differ?”, “How much are they alike?”, and, finally, “What (if anything) should be done as a result of understanding this?”

Practical steps to being sceptical

If you’re a research student, it is well worth reading recommendations about non-biased experimental design and analysis practices in full (Simmons et al., 2011; Rippon et al., 2014). Geoff Cumming’s book about new statistical methods is also well worth reading for a general understanding of why p-values should not be trusted, regardless of your area of study (Cumming, 2012). I hope that these recommendations will be a good starting point for students and non-specialist readers who want a quick guide to being sceptical about reading and researching sex/gender differences. The tips are drawn from recent articles about gender bias in research, and my previous book on modern statistical methods in Human-Computer Interaction (Robertson & Kaptein, 2016b).

It’s important to note that although this article is critical of empirical research and the misuse of statistical methods, I am absolutely not arguing that science is doomed and that we should resort to anecdotal understanding. I’m arguing that if we’re going to use science to investigate sex/gender similarities or differences (if we must), we should use the most robust methods we have.

Next time you read a news report about gender or sex differences, take it with a pinch of salt. Look out for these red flags: the phrases “hardwired”, or other suggestions that sex differences are genetically predetermined or unchanging (Maney, 2016); leaps from animal models or preclinical research to speculation about human behaviour; evolutionary hand-waving explanations; and studies of only small samples of people.

If you spot any of these red flags, adjust your internal scales of belief downwards.

Certainly check it out a bit more before you choose to man- or lady-splain it to your colleagues.

If you have read an article which seems to be well reported, and you are considering using it to inform a gender policy you are working on or including it in an argument for your own research, track down the original paper to visualise the effect sizes using the tool at <https://sexdifference.org/>. Enter the sample size, mean and standard deviation for the experimental groups (e.g. men and women), and the software generates a graph which shows the overlap between the distributions on the independent variable. This gives you an intuitive grasp of how large or small the differences between men and women are in a way that p-values do not. Then ask yourself whether the difference in the independent variable between the groups is enough to make an actionable real world difference. For example, if the independent variable is a reaction time, and an effect size of one millisecond difference is observed, would that millisecond be perceptible (or dangerous) in everyday situations? If the independent variable is an attitude survey, what does it mean if one group tends to answer “strongly agree” rather than “agree” on two questions in a larger set?

If you’re a researcher, and you are planning to look for sex/gender differences, consider: is this important, and why? Are there other issues which might be more important (Hyde, 2014)? Is sex/gender a proxy for other factors such as body mass or levels of particular hormones which might be more informative? Do you have a well-formed theory for why there would be differences? The Gendered Innovation project (<https://genderedinnovations.stanford.edu/researchers.html>) offers useful checklists which can help you to decide which considerations of gender may be important for your area of study (Klinge, 2013).

Research Report Bingo		
Hardwired	Mars	Intuitive
Evolution	Add your own	Testosterone
Mice	Venus	Predetermined

Assuming that you’ve decided to go ahead with a study of gender differences, bear in mind that doing subgroup analysis will reduce your ability to detect an effect, and so you will need a large sample size. If you’re at the stage of sketching out ideas, glance at Cohen’s power primer table which shows roughly

how many participants are required to detect small, medium or large effects with different numbers of comparison groups in the behavioural sciences (Cohen, 1992). You’ll be astonished. For example, consider a study which compares two groups in a between-subjects design. For analysis using a two-tailed independent samples t-test with alpha set at .05, with a power of .80 and attempting to detect a medium-sized effect (Cohen’s $d = .30$), the researcher should recruit 177 participants in each group.

Do a power calculation (you can use R or a free online tool) before proceeding with the experimental design. Statistical power is a function of sample size, population effect size and the significance criteria (known as the alpha value, which is set by convention in behavioural sciences at .05).

Decide on your hypotheses, inclusion criteria and when you will stop collecting data in advance. This will help to prevent “fishing trips” or “p-hacking” later – these are pejorative terms for the practice of running various unplanned analyses until you find the result you wanted, or a significant result you think a journal will publish. You could also consider whether you want to use Null Hypothesis Significance Testing at all. A Bayesian approach might be more constructive because it enables you to estimate the strength of new evidence for different hypotheses based on prior evidence (Kruschke, 2010).

Don't torture your data to make it confess. Look at graphs of your data first. Avoid running squillions of tests without correcting for multiple comparisons. Use tests appropriate to finding interaction effects (such as ANOVA). Check your effect sizes before you start claiming substantial differences (Robertson & Kaptein, 2016a). Remember: “How much of a difference” is usually more interesting than “Does a difference exist?”

When you write up your study, pay attention to fair statistical communication (Dragicevic, 2016). Avoid the temptation to over-interpret your results, or to over-emphasise small differences. Talk your university press officer down from writing a cute press-baiting story about why women and men are from different planets. Insist that what really matters is how your results can help the world, not further divide it by reinforcing stereotypes.