Dr. Jason M. Chin, MA (psychology), PhD (psychology), JD
jason.chin@sydney.edu.au
Lecturer, University of Sydney, School of Law
President, Association for Interdisciplinary Meta-research
and Open Science (AIMOS)

21 June 2021

Australian Law Reform Commission
impartiality@alrc.gov.au
PO Box 12953 – George Street Post Shop
Queensland 4003

RE: Judicial Impartiality Consultation Paper - Submissions

Dear Commissioner and ALRC team:

Thank you for circulating your April 2021 Consultation Paper, 'Judicial Impartiality'. These are my submissions in response to it.

I write from the perspective of a law and psychology researcher and as a researcher who is concerned about the quality and generalisability of the social scientific research relating to cognitive bias. I am also a lecturer at the School of Law at the University of Sydney, where I teach law and psychology, and I am the President of AIMOS (the Association for Interdisciplinary Meta-research and Open Science). AIMOS is an international organisation of over 200 individuals advancing the field of research on research (meta-research) and open science. I do not, however, speak for either of these institutions.

My submission is primarily about Consultation Question 21:

> What further steps, if any, should be taken by the Commonwealth courts or others to ensure that any implicit social biases and a lack of cultural competency do not impact negatively on judicial impartiality, and to build the trust of communities with lower levels of confidence in judicial impartiality? Who should be responsible for implementing these?

However, my submission pertains to many of your other questions and proposals as well. For instance, I will also address your Background Paper JI6 ('Cognitive and Social Biases in Judicial Decision-Making'). I understand that it was not meant to be your definitive position on the research behind cognitive and social biases, and that further research is planned after this consultation stage. Still, I was concerned with much of the research summarised in that report (and I believe many more active researchers in social and cognitive psychology would be as well). I hope my submission will lead you to engage in more systematic research on cognitive and social biases and adopt a more sceptical approach towards that work.

What is the replication crisis and credibility revolution?

Before discussing research on the psychology of bias, it is necessary to briefly review the controversies and reforms going on in psychology and in science more generally. These event and changes, along with the meta-research that has accompanied them, have greatly increased our knowledge about how to assess psychological research.

Led in part by psychological researchers, science is undergoing a massive reform movement sometimes referred to as a Credibility Revolution (Vazire, 2018). This movement was instigated by what has been called a replication crisis (Munafò et al, 2017). In psychology, this began about 10 years ago when researchers started reporting difficulties replicating (i.e., conducting the same study with different subjects) research published in leading journals by eminent authors. One of the most salient and notorious of these was research that studied priming: exposing participants to brief and even subconscious information and observing profound behavioural responses. I note here that you reviewed and seemed to accept priming findings in your Background Report (p 12). Researchers following these protocols exactly as published did not observe any priming effects. Nelson and colleagues (2018) describe these events as follows:

> Doyen et al. (2012) reported a failure to replicate one of the most famous findings in social psychology, that priming people with elderly stereotypes made them walk more slowly (Bargh et al. 1996). This prompted a lively and widely publicized debate, which, in turn, prompted Nobel Prize winner Daniel Kahneman to write a widely circulated email calling for researchers to resolve the debate by conducting systematic replications. (We have archived some of these exchanges at https://osf.io/eygvz/.) Perhaps not coincidentally, replication attempts soon became much more common.

As Nelson and colleagues note, large teams of researchers gathered to test, with much larger samples than the original's used and with more rigorous methods, over a hundred studies published in the most widely cited journals (one project drew only on social science studies published in *Nature* and *Science*, which are very influential journals). As summarised in the table below, in only about 50% of studies was the same effect observed. And, very consistently, even if the effect was observed, it was about half the size of the original (see also Kvarven et al, 2020).

Project	Field	# Replication studies	Statistically significant in the same direction as the original
Estimating the replicability of psychological science	Psychology	97	36%
Evaluating replicability of laboratory experiments in economics	Economics	18	61%
Investigating Variation in Replicability (Many Labs 1)	Psychology	16	88%
Many Labs 2: Investigating Variation in Replicability Across Samples and Settings	Psychology	28	54%
Many Labs 3: Evaluating participant pool quality across the academic semester via replication	Psychology	9	33%
Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015	Social Sciences	21	62%

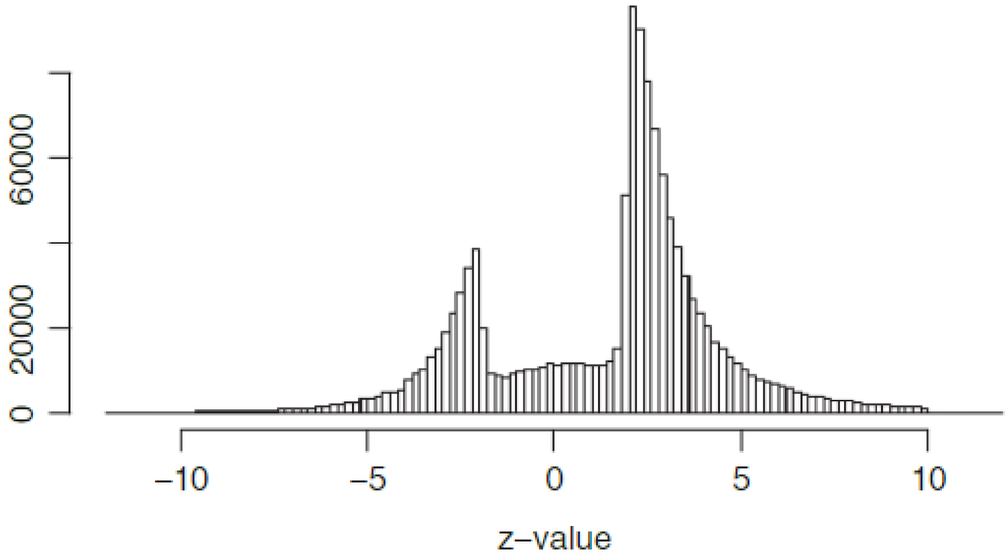
Here, I should note that some of these studies also tried to replicated subconscious priming effects and failed (e.g., two priming studies in Many Labs 1, across approximately 11,000 participants, showed no priming effect). However, other effects you mention in your Background Report are robust. For example, the anchoring effect (bias) you review at page 8 of your report was replicated in four scenarios with over 20,000 participants in Many Labs 1 (Klein et al, 2014).

Still, the generally bleak picture in the above table likely contributed to the results of a survey of 1,576 researchers in *Nature*, in which 52% agreed there was a significant crisis in science, 38% reported there was a slight crisis and only 3% said there was no crisis (Baker, 2016).

What are the contributors to the replication crisis?

Many traditional research practices and norms contributed to the replication crisis. I will review a few relevant to your inquiry and Background Paper, but easy-to-read reviews contain a fuller picture (see Hardwicke et al, 2020).

- **Publication bias** refers to any systematic bias in what is published in research outlets, like the journals and reports you cited in your Background Report. Most relevant to my submission is the file drawer effect, whereby studies that failed to find something (e.g., support for the hypothesis that implicit bias training is effective) are not published. There is a vast amount of evidence that the file drawer effect is rampant across the sciences (see Fanelli, 2012, ‘Negative results are disappearing from most disciplines and countries’). An excellent visual demonstration of the file drawer effect is below. The authors analysed nearly 100,000 studies and produced the below distribution of the results of those studies. They found that Z-values around 0 (i.e., studies in which no difference between the treatment and control was found) are systematically underrepresented in the published literature (the y-axis is the number of published studies with the given Z-score). That stark visual absence of effects around 0 is the file drawer effect.



(from van Zwet and Cator (2020) reproduced under a Creative Commons Attribution-NonCommerical-NoDerivs license).

- **Undisclosed flexibility in the research process** (Beerdsen, 2021) also contributed the replication crisis. For example, researchers in many fields, including psychology (John et al, 2011), regularly exclude outliers after observing their effect on the findings. This biases the results towards finding what the researcher is looking for (Simmons et al, 2012).

- **Small sample sizes** have also contributed to replication crisis (Ioannidis, 2005). Small samples are more susceptible to being influenced by random variation, especially in studies of people (who vary many ways unrelated to what is being studied). For example, some judges randomly assigned to an implicit bias training condition may simply possess fewer biases or be unusually receptive to training (in ways that the average judge is not). This is called sampling error, and it is pronounced with small samples. Large samples, which have traditionally been rare in much of psychology, are needed to account for this variation and provide credible results. Moreover, within the bias training literature, there is a profound effect whereby it is the studies with larger samples that show smaller effects and the studies with fewer participants that show that implicit bias training is very effective (Paluck et al, 2021). This *strongly suggests* that there are many unpublished studies with small sample sizes showing that implicit bias training does not work or does not work very well – and that the literature is substantially overstating the effectiveness of implicit bias training. I discuss this further below.

- **Eminence** (i.e., status bias) can bias published research when editors choose to publish research based on the status of the researcher instead of scrutinizing the quality of the research (Vazire, 2017).

What is the credibility revolution?

In response to the replication crisis, many fields are undergoing a credibility revolution (Vazire, 2018). They are using more transparent methods (e.g., open data, open materials, preregistration, which is publicly describing a study’s methods prior to conducting the study to avoid publication bias and practices like ad hoc data exclusions) so that their work can be critically appraised by community. Over time, by giving other researchers access to traditionally hidden parts of the research process, published results will be more credible.

Implicit biases of judges

You reviewed work measuring judges' implicit biases and whether they affect their self-reported attitudes and decisions. For example, you said this about one study:

> The outcomes were mixed: when judges were specifically told of the race of the defendant there was no difference in outcome; but when they were subliminally primed to think about race, differences did exist, and these differences correlated with IAT scores. The researchers concluded that 'judges harbor the same kinds of implicit biases as others; that these biases can influence their judgment; but that given sufficient motivation, judges can compensate for the influence of these biases.'

Findings like this are not credible enough to guide policy. This study used small sample sizes, a poorly-understood assessment tool and construct (the IAT, see below), unconscious priming (see above), and was conducted during in a time when psychological researchers widely used undisclosed flexibility in their methods to make their results seem more probative than they were (and almost no transparency was required of authors). You should not place any weight on this study.

The Implicit Associations Test (IAT)

I was glad to see that you expressed some reservations about the IAT in some parts of the Background Paper (although not throughout), saying 'It has also been argued that, while IAT scores may be predictive of behaviour when used in the aggregate, they should not be used to predict the behaviour of individuals.'

However, I believe you did not go far enough with these reservations. It is not just that we do not know much about the relationship between the IAT and behaviour (in the aggregate or individually). Rather, we know very little about what construct the IAT measures. For instance, Gawronski (2019) wrote:

> In fact, the growing skepticism has become so pervasive that even early proponents have started to question the explanatory value of implicit bias (e.g., Forscher, Mitamura, Dix, Cox, & Devine, 2017), with some critics dismissing the construct as entirely irrelevant for the psychological understanding of social discrimination (e.g., Blanton & Jaccard, 2017; G. Mitchell, 2018).

Before making any recommendations about measuring judges' biases with the IAT, I recommend you conduct further research and consult with a psychologist who is actively studying it.

Implicit bias training

I also was heartened to see that your Background Paper expressed some caution when reviewing implicit bias training (p 23) and was based on a fairly rigorous and transparent review performed by the UK Equality and Human Rights Commission report (2018). It should be noted that that report (p 9) ‘found that only 18 sources of evidence were both relevant to the research question and adopted the minimum standards for quality research. The number of rigorous studies assessing the effectiveness of UBT is small and this is a significant finding in itself.’

But, it is not just that there are few studies. Rather, unconscious bias training sits in a more general field, prejudice reduction, that exhibits *extreme publication bias*. In a 2021 meta-analysis, Paluck and colleagues found [**emphasis added**]:

> in **every theoretical domain we find unmistakable indications of publication bias**: **Large-N lab, online, or field studies that generate precise results tend to produce much weaker effects than small-N studies** that generate results with large standard errors…
>
> Although our result is robust to the precise manner in which meta-analytic results are calculated, the results are not robust to the most basic assessments of study quality. We offer just one: study size, or number of participants. In the absence of publication bias, we should obtain similar average effect estimates from small and from large studies. However, Table 1 **demonstrates a powerful inverse relationship between study size and effect size**. Restricting attention solely to the quintile of smallest studies (i.e., the 74 studies that allocate 25 or fewer participants to the treatment condition), we obtain a meta-analytic estimate of d = 0.61 (SE = 0.05). This large effect size would on average move a person who feels mildly negatively toward Black people at 40 to a solidly neutral feeling of 53. By contrast, the 73 studies in the highest quintile of study population size, which allocate 78 or more participants to the treatment group, generate a meta-analytic estimate of d = 0.19 (SE = 0.02). These larger studies predict that on average, interventions would change feelings toward Black people in a positive direction but only by approximately 4 points on the scale, such that people who started out feeling 40 would still rate their feelings as mildly negative (44) following an intervention. Studies with intermediate-size treatment groups produce intermediate-size effects. The relationship between our meta-analytic effect size and the size of a study’s treatment group is highly significant ($p < 0.0001$). Importantly, this finding is not an artifact of research methodology—e.g., of online studies with larger samples finding weak effects…

I note that the UK review did not take into account publication bias, as other systematic review formats would.

Furthermore, most studies that test unconscious bias training do not take steps to control for researcher bias (through, e.g., using more transparent methods) (Paluck, 2021, p 539).

Who should be involved in law reform?

Finally, my critiques above raise the question of who should be involved in this project and in law reform projects generally.

For this project, it is notable that the advisory panel does not contain any psychologists, let alone a social or cognitive psychologist who studies bias or implicit bias. Most of the expertise is in law (there is a psychiatrist, but that field focuses on treating patients and would not be involved in any of the research covered in the Background Report I have been commenting on).

In these circumstances, Naomi Oreskes' (a philosopher of science) work on diversity seems particularly apt (2019, p 53): 'The greater the diversity and openness of a community and the stronger its protocols for supporting free and open debate, the greater the degree of objectivity it may be able to achieve as individual biases and background assumptions are "outed," as it were, by the community.'

Just as you are studying the biases of judges, you might want to consider lessening the biases of an advisory committee filled with individuals from the legal profession. You can do so by improving its disciplinary diversity. Along these lines, I highly suggest you add to the advisory panel a social or cognitive psychologist that is actively studying bias.

Moreover, I suggest you begin working with meta-researchers, a field in which Australia is leading the world in. They can provide valuable insights about evaluating research. For instance, you may wish to contact Professor Fiona Fidler (University of Melbourne), who is leading a large DARPA-funded (US Defense Advanced Research Projects Agency) project aimed at determining what social scientific research is credible enough for DARPA to rely on and what research they should not rely on. Just as it is important that a defence agency base its policies on reliable social science, so should law reform bodies. Professor Simine Vazire (University of Melbourne, but currently based in Sydney) would also be an excellent person to work with or to speak at your commission.

Along these lines, Brian Nosek (2019), who is one of the leaders of the meta-research field, provided the following testimony to the U.S. House of Representatives about meta-research, uncertainty, the DAPRA project, and policy-making [**emphasis added**]:

> In policymaking, it is important to use the best available evidence for rulemaking. **There will always be occasions in which the best available evidence is not fully transparent or has unknown reproducibility. Using the best available**

> **evidence does not mean using it blindly or overconfidently. There are many factors that affect the quality of research, the certainty of its conclusions, and its generalizability to the policy context. Explicitly representing the uncertainty of evidence will help policymakers make better decisions.** When the evidence is more uncertain, policymakers could ensure that implementation of the policy includes mechanisms to evaluate its success. And, by knowing the uncertainty of evidence, policymakers could direct resources to supporting research to address those certainty gaps and improve the overall evidence base. For example, DARPA's SCORE program is investigating whether machine algorithms could automatically assess the credibility of research claims. If successful, this could provide an initial filter to inform the translation of research evidence into practice, and prioritization of research funding to topics of national and research interest.

Until tools like DARPA's are widely available, I believe law reform commissions like the ALRC should take advantage of the wealth of meta-research expertise in Australia and beyond, as well as AIMOS as an institution. Beyond the names above, I would be happy to provide further connections (e.g., Professor Adrian Barnett at QUT, Professor Shinichi Nakagwa at UNSW, Dr. Matthew Page at Monash).

You may also wish to begin conducting or funding systematic reviews (e.g., the UK Equality and Human Rights Commission report was a systematic review). These reports are transparent about how they will search the literature and report studies they include and exclude to avoid bias. They also provide mechanisms for assessing quality and publication bias.

Thank you for your consideration. I would be happy to discuss this further.

Sincerely,

Jason Chin

References

Baker M (2016) 1,500 scientists lift the lid on reproducibility: Nature News & Comment. *Nature*. https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970.

Beerdsen E (2020) Litigation Science after the Knowledge Crisis. *Cornell Law Review*.

Camerer, CF et al (2016) Evaluating replicability of laboratory experiments in economics. 351 *Science* 1433. https://doi.org/10.1126/science.aaf0918 .

Camerer, CF et al (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 637. https://doi.org/10.1038/s41562-018-0399-z.

Ebersole C, et al (2016) Many Labs 3: Evaluating participant pool quality across the academic semester via replication. 67 *Journal of Experimental Social Psychology* 68. https://doi.org/10.1016/j.jesp.2015.10.012.

Equality and Human Rights Commission (2018) Unconscious bias training: an assessment of the evidence for effectiveness. https://www.equalityhumanrights.com/en/publication-download/unconscious-bias-training-assessment-evidence-effectiveness.

Fanelli D (2012) Negative results are disappearing from most disciplines and countries. 90 *Scientometrics* 891. https://doi.org/10.1007/s11192-011-0494-7.

Gawronski B (2019) Six Lessons for a Cogent Science of Implicit Bias and Its Criticism. 14(4) Perspectives on Psychological Science 574. https://doi.org/10.1177%2F1745691619826015.

Hardwicke TE, et al (2020) Calibrating the Scientific Ecosystem Through Meta-Research. 7 *Annual Review of Statistics and its Application* 11. https://doi.org/10.1146/annurev-statistics-031219-041104.

Ioannidis JPA (2005) Why Most Published Research Findings Are False. 2(8) *PloS Medicine* E124. https://doi.org/10.1371/journal.pmed.0020124.

John LK, et al (2012) Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. 23(5) *Psychological Science* 524. https://doi.org/10.1177%2F0956797611430953.

Klein RA, et al (2014) Investigating variation in replicability: A ‘many labs’ replication project. 45(3) *Social Psychology* 142. https://psycnet.apa.org/doi/10.1027/1864-9335%2Fa000373.

Klein RA, et al (2018) Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. 1 *Advances in Methods and Practices in Psychological Science* 443. https://doi.org/10.1177%2F2515245918810225.
Kvarven A, et al (2020) Comparing meta-analyses and preregistered multiple-laboratory replication projects 4 *Nature Human Behaviour* 423. https://doi.org/10.1038/s41562-019-0787-z.

Munafò MR, et al (2017) A manifesto for reproducible science. 1(1) *Nature Human Behaviour* 1. https://doi.org/10.1038/s41562-016-0021.

Nelson LD, et al (2018) Psychology’s Renaissance. 69 *Annual Review of Psychology* 511. https://doi.org/10.1146/annurev-psych-122216-011836.

Nosek, Brian (2019) Testimony of Brian A. Nosek, Ph.D. Executive Director Center for Open Science Before the Committee on Science, Space, and Technology U.S. House of Representatives. https://osf.io/preprints/metaarxiv/ve83q/download.

Open Science Collaboration (2015) Estimating the Reproducibility of Psychological Science. 349(6251) *Science* 943. https://doi.org/10.1126/science.aac4716.

Oreskes N, *Why Trust Science?* (Princeton University Press, 2019).

Paluck EL, et al (2021) Prejudice Reduction: Progress and Challenges. 72 *Annual Review of Psychology* 533. https://doi.org/10.1146/annurev-psych-071620-030619.

Simmons JP, et al (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. 22(11) *Psychological Science* 1359. https://doi.org/10.1177%2F0956797611417632.

Van Zwet EW & Cator EA (2020) The significance filter, the winner's curse and the need to shrink. *Statistica Neerlandica*. https://doi.org/10.1111/stan.12241.

Vazire S (2017) Our obsession with eminence warps research 547 *Nature* 7. https://doi.org/10.1038/547007a.

Vazire S (2018) Implications of the Credibility Revolution for Productivity, Creativity, and Progress. 13(4) *Perspectives on Psychological Science* 411. https://doi.org/10.1177%2F1745691617751884.