Manuscript in press at Nature Human Behaviour.

Use caution when applying behavioural science to policy

Authors: Hans IJzerman[1,2], Neil A. Lewis Jr.[3], Andrew K. Przybylski[4], Netta Weinstein[5], Lisa DeBruine[6], Stuart J. Ritchie[7], Simine Vazire[8,9], Patrick S. Forscher[1], Richard D. Morey[5], James D. Ivory[10], Farid Anvari[11]

Institutes: [1]Université Grenoble Alpes, [2]Institut Universitaire de France, [3]Cornell University, [4]University of Oxford, [5]Cardiff University, [6]University of Glasgow, [7]King's College London, [8]University of California, Davis, [9]University of Melbourne, [10]Virginia Polytechnic Institute and State University, [11]University of Southern Denmark

Author Note: Przybylski, Lewis Jr., and IJzerman contributed equally and share first authorship.

Standfirst

Social and behavioural scientists have attempted to speak to the COVID-19 crisis. But is behavioural research on COVID-19 suitable for making policy decisions? We offer a taxonomy that lets our science advance in *Evidence Readiness Levels* to be suitable for policy. We caution practitioners to take extreme care translating our findings to applications.

Researchers in the social and behavioural sciences periodically debate whether their research should be used to address pressing issues in society. To provide a few examples, in the 1940s psychologists discussed using research to address problems related to intergroup relations, problems brought to the fore by the Holocaust and other acts of rampant prejudice. In the 1990s, psychologists debated whether their research should inform legal decision-making. In the 2010s, psychologists argued for advising branches of government as economists often do. And now, in 2020, psychologists and other social and behavioural scientists are arguing that our research should inform the response to the new coronavirus disease (henceforth COVID-19)[1,2].

We are a team mostly consisting of empirical psychologists who conduct research on basic, applied, and meta-scientific processes. We believe that scientists should apply their creativity, efforts, and talents to serve our society, especially during crises. However, the way that social and behavioural science research is often conducted makes it difficult to know whether our efforts will do more good than harm. We will provide some examples from the field of social-personality psychology, where most of us were trained, to illustrate our concerns. This focus is not meant to imply our field alone suffers from the issues we will discuss. Instead, a growing meta-science literature suggests that many other social and behavioural disciplines have encountered similar dynamics as our field.

What are those dynamics? First, study samples, mainly students, are drawn from populations that are in Western (mostly US), Educated, Industrialized, Rich, and Democratic societies[3]. Second, even with this narrow slice of population, the effects in published papers are not estimated with precision, sometimes barely ruling out trivially

small effects under ostensibly controlled conditions. Third, many studies use a narrow range of stimuli and do not test for stimulus generalisability[4]. Fourth, many studies examine effects on measures, such as self-report scales, that are infrequently validated or linked to behaviour, much less policy relevant outcomes[5]. Fifth, independently replicated findings, even under ideal circumstances, are rare. Finally, our studies often fail to account for deeper cultural, historical, political, and structural factors that play important moderating roles during the process of translation from basic findings to application. Together, these issues produce empirical insights that are more heterogeneous than might be apparent from a scan of the published literature.

Confident applications of social and behavioural science findings, then, require first and foremost an assessment of the evidence quality and weighing heterogeneity and the tradeoffs and opportunity costs that follow. We must identify reliable findings that *can* be applied, have been investigated in the world's nations where the application is intended for, and are derived from investigations using diverse stimuli. But the assessment of how "ready" the intervention is *must* be included when persuading decision-makers to apply social and behavioural science evidence, particularly in crisis situations when lives are at stake and resources are limited. Not doing so can have disastrous consequences.

Here we propose one approach for assessing the quality of evidence prior to application and dissemination. Specifically, we draw inspiration from the National Aeronautics and Space Administration's *Technology Readiness Levels* (TRL[6]), a benchmarking system for systematically evaluating the quality of scientific evidence and which has been utilized by the European Commission to judge how ready scientific

applications beyond space flight are for operational environments. TRLs rank a technology's readiness for application from 1 to 9 (see Figure 1). At TRL1, basic principles have been *reliably* observed, reported, and translated to a formal model. In TRL2, basic principles have been developed and tested in an application area. It is not until TRL4, when a prototype is developed, that tests are run in various environments that are as representative of the eventual application area(s) as possible. Later at TRL6, the system is tested in a "real" environment (like ground-to-space). At the very highest level (TRL9), the system has been "flight-proven" through successful mission operations. These TRLs provide a useful framework to jumpstart conversations about how to assess the readiness of social and behavioural science evidence for application and dissemination.

Introducing Evidence Readiness Levels

The desire to "directly inform policy and individual and collective behaviour in response to the pandemic" (p. 461)[1] overlooks existing evidence frameworks and the challenges we identify, illustrating that a simple taxonomy is necessary to have at hand during crises. As a very preliminary step to this end we propose a social and behavioural science variant of TRLs, *Evidence Readiness Levels* (ERLs; Figure 2).

There are several frameworks for assessing evidence quality across different scientific fields. The one that comes closest to what we envision is the Society for Prevention standards for prevention interventions[7], as it incorporates standards for efficacy and dissemination and feedback loops from crisis to theory. However, none of the existing frameworks capture the meta-scientific insights generated in our field in the last decade.

Our ERLs do not map perfectly onto NASA's TRLs and we should not expect them to; there are many differences between behavioural and rocket science. In the social and behavioural sciences we think this process should start with defining problem(s) in collaboration with the stakeholders most likely to implement the interventions (ERL1). These concepts can then be further developed in consultation with people in the target settings to gather preliminary information about how settings or context might alter processes (ERL2). From there, researchers can conduct systematic reviews and other meta-syntheses to select evidence that could potentially be applied (ERL3). These systematic reviews require a number of bias-detection techniques. It is well-known that the behavioural sciences suffer from publication bias and other practices that compromise the integrity of research evidence. Some findings *may* be reliable, but the onus is on us to identify which are and which are not and which generalize or not. Yet, still then, these systematic reviews must be done with an awareness that the currently available statistical techniques do not completely correct for bias and that the resultant findings are at most at ERL3.

Following this, one can gather information about stimulus and measurement validity and equivalence for application in the target setting (ERL4). After, researchers - in consultation with local experts - should consider the potential benefits and harms associated with applying potential solutions (ERL5), and generate estimates of effects in a pilot sample (ERL6). With preliminary effects in hand, the team can then begin to test for heterogeneity in low-stakes (ERL7) and higher-stakes (ERL8) samples and settings which would build the confidence necessary to apply the findings in the real target setting or crisis situation (ERL9).

Even at ERL9, evidence evaluation continues; applications of social and behavioural work, particularly in a crisis, should be iterative so high-quality evidence is fed back to evaluate the effectiveness of the intervention, critical and flexible improvements. Feedback should be grounded in collaboration between basic and applied researchers, as well as with stakeholders to ensure that the resulting evidence is relevant and actionable. Failure to continually re-evaluate interventions in light of new data could lead to unnecessary harm, where even the best evidence was inadequate to predict the intervention's real-world effects.

A benchmarking system such as the ERL requires us to think carefully about the nature of our research that can be applied credibly and guides where research investments should be made. For example, we can better recognise that our goal of gathering reliable insights (ERL3) provides a necessary foundation for further collective efforts that scaffold towards scalable and generalisable interventions (ERL7). Community experts, identifying relevant theories, and extensive observations are key to framing challenges and working with interdisciplinary teams to address them (ERL1). Behavioural scientists from different cultures then discuss how interventions may need to differ in nature across context and cultures. The multidisciplinary and multi-stakeholder nature of ERLs require us to fundamentally rethink how we produce, and communicate confidence in, application-ready findings.

The current crisis provides a chance for social and behavioural scientists to question how we understand and communicate the value of our scientific models in terms of ERLs. It also requires us to communicate those ERLs to policy-makers so that they know whether we are making educated guesses (ERL3 or below) or can be

confident about the application of our findings because we have tested and replicated them in representative environments (ERL7). When providing policy advice on the basis of scientific evidence, it is important to understand and be able to explain whether and how recommendations would impact relevant people under a range of circumstances that are highly relevant to the crisis in question (ERL7).

Even if findings are at ERL3 after having assessed evidence quality of primary studies, we have little way of knowing how much positive, or unintended negative, consequences an intervention might have when applied to a new situation. We are concerned to see social and behavioural scientists making confident claims about the utility of scientific findings for solving COVID-19 problems without regard for whether those findings are based on the kind of scientific methods that would move them up the ERL ladder[1]. The absence of recognised benchmarking systems makes this challenging. While it is tempting to instead qualify uncertainty by using non-committal language about the possible utility of existing findings (e.g., “may”, “could”), this approach is fundamentally flawed because public conversations generally ignore these rhetorical caveats[8]. Scientists should actively communicate uncertainty, particularly when speaking to crises. Communicating that their ERL is only at 3 or 4 would empower policy makers how to weigh our advice in terms of their options. Reaching a higher ERL is extremely complicated and will require radical changes in the way we conduct research, also beyond crises.

How Social and Behavioural Scientists Can Advance their ERLs

The field of genetics started in a position similar to the position that many behavioural sciences find themselves in now with small, independently collected

samples that produced unreliable findings. Attempts to identify candidate genes for many constructs of interest kept stalling at TRL1/ERL4. In one prominent example, 52 patients provided genetic material for an analysis of the relationship between the *5-HTT* gene and major depression[9], a finding that spurred enormous interest in the biological mechanisms underlying depression. Unfortunately, as with the current situation in psychology, these early results were contradicted by failed replication studies[10].

Technological advances in genotyping unlocked different approaches for geneticists. Instead of working in isolated teams, geneticists pooled resources via consortium studies and thereby accelerated scientific progress and quality. Their recent studies (with samples that sometimes exceed 1,000,000) dwarf previous candidate gene studies in terms of sample size[11]. To accomplish this, geneticists devoted considerable time to developing research workflows, data harmonization systems, and processes that increased the accuracy of their measurements. The new methodologies are not without flaws: for example, there is substantial scope for expanding the representativeness of study cohorts. But the progress that consortium research in genetics has made in a short time is impressive.

In recent years we have observed similar progress in the psychological sciences going from single, small-sample studies to large-scale replications[12,13] and novel studies[14] to the building of the prerequisite infrastructure to facilitate team science. One example is the Psychological Science Accelerator (PSA), a large, standing network with experts facilitating study selection, data management, ethics, and translation[15]. While the PSA is making important progress, problems surrounding measurement validity, sample generalizability, and organizational diversity (40% of its leadership is from North

America) which affects the networks ability to accurately interpret findings, still present material challenges to the applicability of their projects. Therefore, the PSA will require substantial improvement and investment before it can generate practical ERL7-level evidence and further develop our proposed framework.

The COVID-19 crisis underscores the critical need to bring the social and behavioural sciences in line with other mature sciences. Diverse consortia of researchers with expertise in philosophy, ethics, statistics, and data and code management are needed to produce the kind of research required to better understand people the world over. Realising this mature, inclusive, and efficient model necessitates a shift in the knowledge production and evaluation models that guide the social and behavioural sciences.

Be cautious when applying social and behavioural science to policy

On balance, we hold the view that the social and behavioural sciences have the *potential* to help us better understand our world. However, we are less sanguine about whether many areas of social and behavioural sciences are mature enough to provide such understanding, particularly when considering life-and-death issues like a pandemic. We believe that, rather than appealing to policy-makers to recognise our value, we should focus on earning the credibility that legitimates a seat at the policy table. The ERL taxonomy is a sample roadmap for achieving this level of maturity as a science, and for accurately and honestly communicating our current state of evidence. Collaborations among large and diverse teams with local knowledge and multi-disciplinary expertise can help us move up the evidence ladder. Equally important, studies in the behavioural sciences must be designed to move up this ladder

incrementally. Designing an ERL6 study that is built on a shaky ERL1 foundation will be of little use. Moving up requires investment, thought, and, most important of all, epistemic humility. Without a systematic and iterative research framework, we believe that behavioural scientists should carefully consider whether well-intentioned advice may do more harm than good.

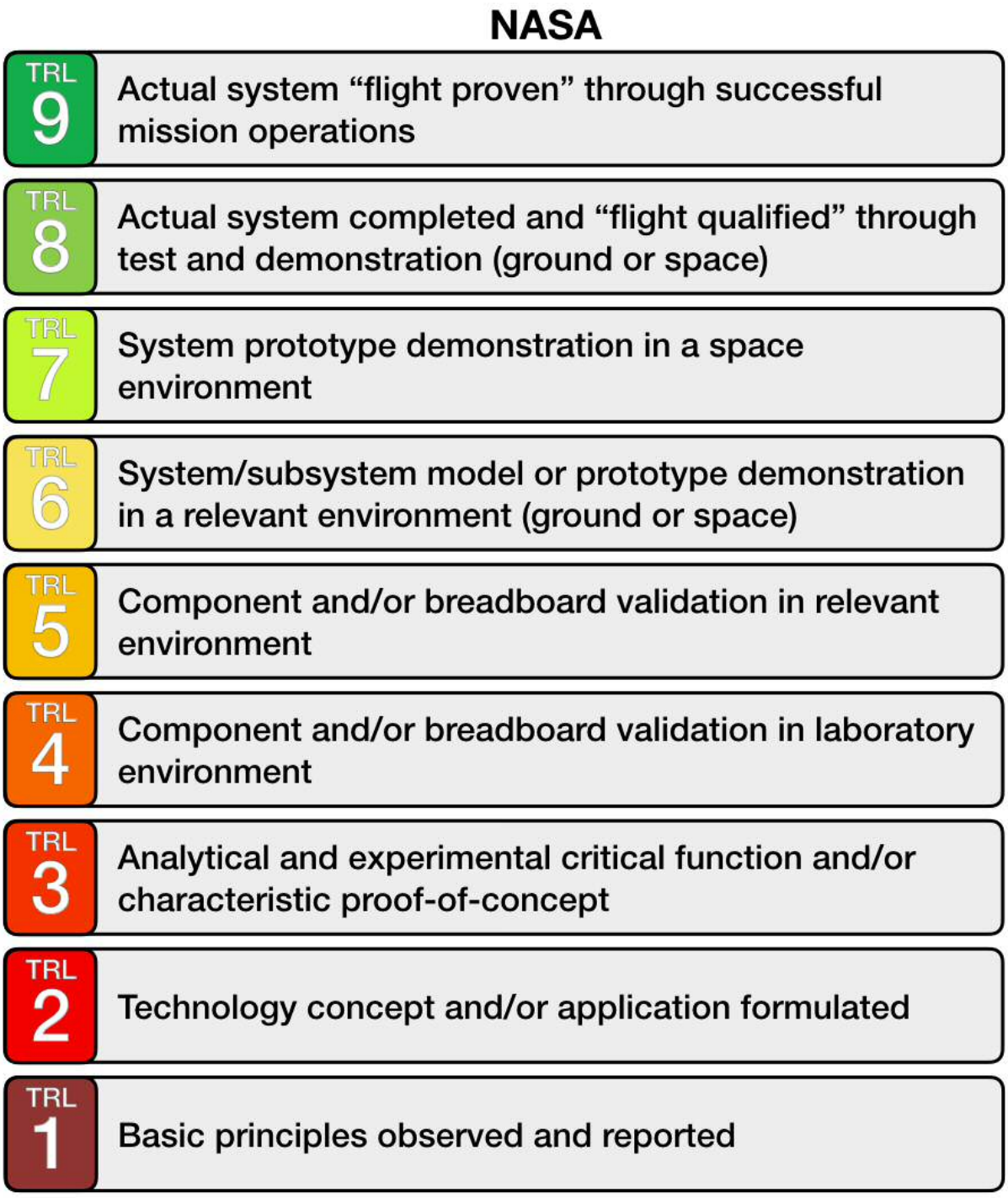

Figure 1. NASA Technology Readiness Levels. Original figure source: https://www.nasa.gov/directorates/heo/scan/engineering/technology/txt_accordion1.html

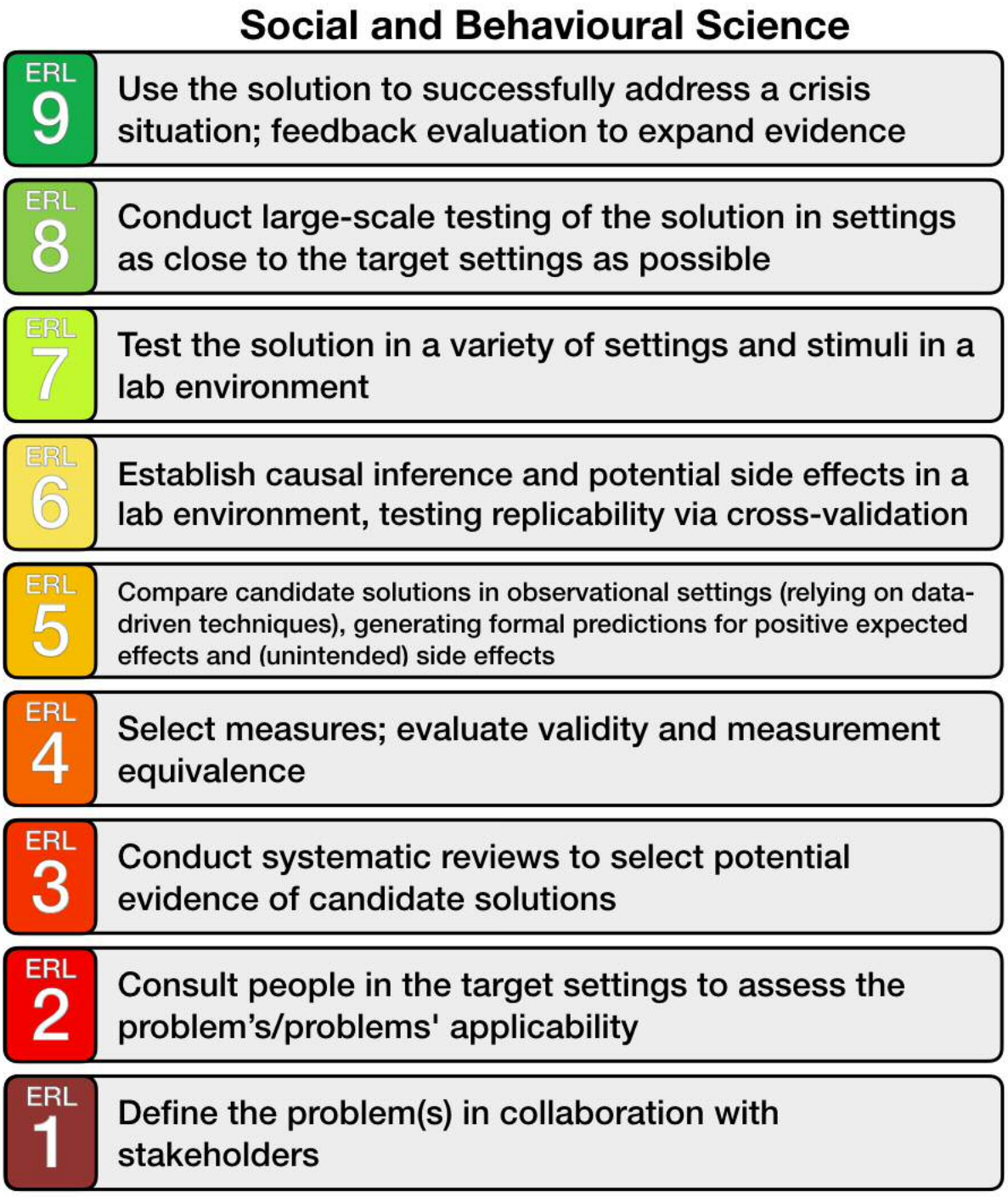

Figure 2. Proposed Social and Behavioural Sciences Evidence Readiness Levels

References

(1) Van Bavel, J. J., Boggio, P., Capraro, V., Cichocka, A., Cikara, M., Crockett, M., ... & Ellemers, N. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behavior, 4,* 460–471 .

(2) Syed, M. (2020). Psychology of COVID-19 preprint tracker. *Available at https://bit.ly/3eykL1s*.

(3) Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

(4) Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54-69.

(5) Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin*, *132*(2), 249.

(6) NASA (2012). Technology Readiness Level. *Available at https://go.nasa.gov/2XKbFsq.*

(7) Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science*, *16*(7), 893-926.

(8) Adams, R. C., Sumner, P., Vivian-Griffiths, S., Barrington, A., Williams, A., Boivin, J., ... & Bott, L. (2017). How readers understand causal and correlational expressions used in news headlines. *Journal of Experimental Psychology: Applied, 23*(1), 1.

(9) Heils, A., Teufel, A., Petri, S., Stöber, G., Riederer, P., Bengel, D., & Lesch, K. P. (1996). Allelic variation of human serotonin transporter gene expression. *Journal of Neurochemistry*, *66*(6), 2621-2624.

(10) Gillespie, N. A., Whitfield, J. B., Williams, B. E. N., Heath, A. C., & Martin, N. G. (2005). The relationship between stressful life events, the serotonin

transporter (5-HTTLPR) genotype and major depression. *Psychological Medicine*, *35*(1), 101-111.

(11) Jansen, P. R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A. R., ... & Savage, J. E. (2019). Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nature Genetics*, *51*(3), 394-403.

(12) Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R.B., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142-152.

(13) Grahe, J. E., Brandt, M. J.,IJzerman, H., Legate, N., Wagge, J., Weisberg, Y. J., & Wiggins, B. J. (2016). Collaborative Replications and Education Project (CREP). *Available at osf.io/wfc6u.*

(14) IJzerman, H., Lindenberg, S., Dalğar, İ., Weissgerber, S.S., Vergara, R. C., Cairo, A. H., ... & Hall, C. (2018). The Human Penguin Project: Climate, social integration, and core body temperature. *Collabra: Psychology*, *4*(1).

(15) Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... & Castille, C. M. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501-515.

Acknowledgements: The preparation of this work was partly funded by a French National Research Agency ”Investissements d’avenir” program grant (ANR-15-IDEX-02) awarded to HI, a Huo Family Foundation grant to AKP, an ERC 647910 (KINSHIP) grant awarded to LD, and an ERC 851890 (SOAR) grant awarded to NW.

Competing interests: PSF, HI, and NAL are on the board of directors of the Psychological Science Accelerator network referenced in the manuscript. The remaining authors declare no competing interests.