Northumbria
University
NEWCASTLE

Evolutionary Psychological Science
https://doi.org/10.1007/s40806-019-00192-2

METHODOLOGY ARTICLE

How Diverse Are the Samples Used in the Journals '*Evolution & Human Behavior*' and '*Evolutionary Psychology*'?

Thomas V. Pollet[1] · Tamsin K. Saxton[1]

Abstract
Psychologists regularly draw inferences about populations based on data from small samples of people, and so have long been interested in how well those samples generalise to wider populations. There is a consensus that psychology probably relies too much on samples from Western, Educated, Industrialised, Rich and Democratic (WEIRD) societies and among those from university students. Online surveys might be used to increase sample diversity, although online sampling still reaches only a restricted range of participants. Studies from evolutionary psychology often seek to uncover aspects of evolved universal characteristics, and so might demonstrate a particular interest in the use of diverse samples. Here, we empirically examine the samples used in the 2015–2016 volumes of '*Evolution & Human Behavior*' (104 articles) and '*Evolutionary Psychology*' (76 articles). Our database consists of 311 samples of humans (median sample size = 186). The majority of samples were either online or student samples (70% of samples), followed by other adult Western samples (19%). Two hundred fifty-three (81%) of the samples were classified as 'Western' (Europe/North America/Australia). The remaining samples were predominantly from Asia ($N = 37$; 12%, mostly Japan). Only a small fraction of the samples were taken from Latin American and Caribbean ($N = 8$) or African ($N = 6$) countries. The median sample size did not differ significantly between continents, but online samples (both paid and unpaid) were typically larger than samples sourced offline. It seems that the samples used are more diverse than those that have been reported in reviews of the literature from social and developmental psychology, perhaps because evolutionary psychology has a greater inherent need to test hypotheses about an evolved and universal human nature. However, it is also apparent that the majority of samples within contemporary evolutionary psychology research remain WEIRD.

Keywords Cross-cultural samples · WEIRD · Sampling · Participants

A recurrent criticism of psychology as a science is the lack of diversity. This lack of diversity refers not solely to the producers of psychological science (e.g. Adair et al. 2002; Bauserman 1997; Cole 2006) or to the topics studied (e.g. Berry 2013), but also to the samples that are used as the basis for drawing inferences. This criticism is recurrent and has been voiced approximately once a decade since around 1965 (e.g. Arnett 2008; Gallander Wintre et al. 2001; Graham 1992; Henrich et al. 2010; Henry 2008; Nielsen et al. 2017; Schultz 1969; Smart 1966). Following a review of the participants in psychological studies, Schultz (1969:218) wrote: '*The extremely small percentage of studies sampling the general adult population was particularly disturbing; none of the studies published in the Journal of Experimental Psychology during those years used a sample of the general population*'.

There is an indisputable geographical bias to the populations sampled by psychologists. For example, a review of articles published between 2006 and 2010 in the three experimental developmental psychology journals with the highest impact factors found that over 90% of the research participants came from Australia, Canada, Europe, the USA, or New Zealand, while under 3% of the participants in the research studies were from Africa, Asia, Central and South America and the Middle East and Israel (Nielsen et al. 2017). Similarly, in the flagship journal '*Journal of Personality and Social Psychology*', 96% of the papers published in 2012 were based on WEIRD samples (Kurzban 2013). Further, from its inception, psychology has relied heavily on undergraduate samples, a situation that has not changed substantially over time. For

✉ Thomas V. Pollet
thomas.pollet@northumbria.ac.uk

[1] Department of Psychology, Northumbria University, NB 165, Northumberland Building, 2 Ellison Place, Newcastle upon Tyne NE1 8ST, UK

Published online: 12 March 2019

example, Gallander Wintre et al. (2001) reviewed 1179 articles spanning six journals across the different subdivisions of psychology and found 68% of the samples to be student samples. They also found that, if anything, the reliance on student samples had increased between 1975 and 1995. A classic paper by Sears (1986) reviewed papers published in 1980 in three mainstream social psychology journals and found that 82% of the samples used students in some form, and 75% used undergraduate students (mainly from the USA) exclusively. Likewise, the 1995 editions of two leading social psychology journals ('*Journal of Experimental Social Psychology*' and '*Journal of Personality and Social Psychology*') used undergraduate students as participants in 95.8% and 70.6% of all cases respectively (Gallander Wintre et al. 2001), and Arnett (2008) calculated that 74% of the samples in the journal '*Social Psychological and Personality Science*' were from student populations. The issues relating to sampling are not limited to (social) psychology, and similar concerns have been voiced in other related disciplines such as consumer research (Peterson 2001), education research (Usher 2018), behavioural economics (Levitt and List 2007) and business research (Bello et al. 2009). For example, Peterson (2001) reviewed the literature in consumer research and found 86% of the samples to be from students.

Online Participants

Perhaps in part as a response to these sorts of criticisms, (social) psychologists have increasingly turned to online platforms to recruit participants who are not students (e.g. Gosling et al. 2004; Gosling et al. 2010; Gosling and Johnson 2010). Over the past decade, there has been a strong increase in the use of participants crowdsourced via online platforms, such as Amazon MTurk or CrowdFlower (e.g. Buhrmester et al. 2011; Paolacci and Chandler 2014). Such expansion has benefited psychological research in many ways. For example, results from classic behavioural experiments (e.g. the Stroop Task; Stroop 1935) were shown to replicate well on these online platforms (Crump et al. 2013). Even when studying political ideologies, it seems that MTurk is well-suited (Clifford et al. 2015). However, although Amazon boasts > .5 million participants, the actual pools from which participants are sampled are much smaller, with the population estimated at around 7300 individuals (Bohannon 2016). Moreover, while crowdsourcing platforms such as Amazon's MTurk allow for sampling more diversely than typical student samples, for example with respect to age range, the participants remain predominantly WEIRD, with some notable exceptions (e.g. Raihani et al. 2013).

Evolutionary Psychology as an Exception?

Since its inception, evolutionary psychology has stressed the importance of human universals (e.g. Buss 1989, 1994, 1995; Cosmides and Tooby 1997; Tooby and Cosmides 1990). Empirical examples where researchers have evaluated whether universals exist by testing across different populations include studies of homicide (Daly and Wilson 1988), economic behaviour (e.g. Henrich et al. 2005) and mate preferences (e.g. Buss et al. 2000; Buss 1989; Schmitt 2005; Shackelford et al. 2005). Cross-cultural universals (see for example those listed in Brown 1991, 2000) are often used by evolutionary psychologists as evidence for adaptive psychological mechanisms (e.g. Buss 1995). It would thus seem, as Apicella and Barrett (2016: p. 92) have argued, that '*perhaps no field of psychology is more strongly motivated and better equipped than evolutionary psychology to respond to the recent call for psychologists to expand their empirical base beyond WEIRD (Western Educated Industrialized Rich Democratic) samples*'. Similarly, Kurzban (2013) argued on the Evolutionary Psychology blog that '*adding evolution to psychology makes the science less WEIRD*'. He found that for the 2012 volume, 65% of the articles in the journal '*Evolution & Human Behavior*' were WEIRD, which contrasts favourably with data for other fields as cited above. This initial evidence suggests that evolutionary psychology is indeed less WEIRD than some subdivisions of psychology.

Here, we examine the samples used in two leading evolutionary psychology journals in more depth. We have no explicit hypotheses, but rather describe the samples used in these two journals according to their geographical origin, age group (adult or child), student status and source (online vs. offline). In addition, we test whether the sample sizes vary based on these categories. Our aim is to provide an up-to-date snapshot of contemporary evolutionary psychology sampling practice, while responding to calls to increase description within science (Scott-Phillips 2018); it is easier to move forward if we better know where we currently stand.

Methods

Coding

As part of a larger project, data relevant to our research questions were captured from all of the articles published in 2015 and 2016 within the journals '*Evolutionary Psychology*' (EP, published by Sage) and '*Evolution & Human Behavior*' (E&HB, published by Elsevier). The 2015 articles were coded by eight coders under the supervision of the first author, and the 2016 articles were coded by the first author. Of course, many key papers on evolutionary psychology are published in other outlets, such as '*Journal of Personality and Social*

Psychology' (e.g. Buss and Shackelford 1997; Kenrick et al. 1995) or '*Psychological Science*' (e.g. Buss et al. 1992). However, it is reasonable to assume that any article published in EP and E&HB is allied to the discipline of evolutionary psychology, broadly conceived. In addition, the choice follows Kurzban's (2013) selection of E&HB for analysis and the publication of his analysis within the Evolutionary Psychology blog, on the website of the eponymous journal.

The coders recorded the geographical region from which the data originated based on the M49 UNDP codes (United Nations 2013: Africa, North America, Latin America and the Caribbean, Asia, Australia, Europe and Oceania (excluding Australia)). If a paper listed more than five countries, we labelled it 'cross-cultural'; we coded each sample individually for papers with one to four geographical samples. We acknowledge that some papers could have an explicitly 'cross-cultural' goal even with just two samples, for example studies establishing measurement invariance. However, our focus here is on the samples being used not the paper, as we believed this to be easier to assess.

After piloting, we settled on the following eight categories for the sample participants: online (paid crowdsourced, such as a sample recruited via MTurk); online (unpaid crowdsourced, such as a sample recruited via Facebook or Twitter); offline (western child); offline (western student); offline (western non-student adult); offline (non-western child); offline (non-western student); offline (non-western, non-student adult). Online studies were subdivided only into paid and unpaid samples given the focus of our research, together with the difficulty of confirming online participant age, student/non-student status and location. Samples from Europe, Australia, New Zealand and North America were coded as Western, while samples from other countries coded as non-Western, following (Stulp et al. 2017). Where disagreement between coders existed, this was resolved via discussion.

Data Analysis

We used R (3.5.1, R Development Core Team 2008) and, among other packages, the R packages *bindrcpp* (0.2.2, Müller 2017), *broom* (0.5.0, Robinson 2017), *dplyr* (0.7.6, Wickham and Francois 2017), *ggplot2* (3.3.0, Wickham 2009), *knitr* (1.17, Xie 2015), *papaja* (0.1.0.9709, Aust and Barth 2016), *plyr* (1.8.4, Wickham and Wickham 2017), *readxl* (1.010, Wickham and Bryan 2017), *stargazer* (5.2.2, Hlavac 2014) and *tidyr* (0.8.1, Wickham 2014) for our analyses. To compare sample sizes, we relied on non-parametric statistics (Siegel and Castellan 1988), with post-hoc comparisons adjusted for multiple testing (Benjamini and Hochberg 1995), given that visual inspection showed that the data were non-normally distributed. We used logarithmic transformations when presenting figures on sample sizes because the largest samples are so much greater than the smallest (Keene 1995). The data and analysis document, including a list of all R packages used in the analysis, are available from the Open Science Framework (http://osf.io/pajhy).

Results

Descriptive Statistics

There were 219 papers, of which 180 papers contained codable samples (EP 76; E&HB 104). Thirty-nine papers could not be coded because they consisted of, for instance, mathematical models, work on non-humans, or literature reviews. Within the 180 codable papers, there were 311 samples, and the median number of samples per paper was 1. The mean sample size was 4094 but this was driven by one extremely large sample (N = 927,134). The median sample size was 186 but sample sizes varied substantially (minimum 11; first quartile 96.5; third quartile 334.5; maximum 927,134).

Figure 1 shows the distribution of samples by geographical region. The majority of samples were from North America (153), followed by Europe (93) and Asia (37). Of the Asian samples, the majority were from Japan (11), followed by China (7) and Israel (6). There were only 6 samples from Africa (4 from Tanzania, 1 from Namibia, 1 from Nigeria) and only 8 from Latin America and the Caribbean (2 from Guatemala; 2 from Curaçao; 2 from Bolivia; 1 community from Northern-Brazil/Southern Guyana/East Venezuela and 1 undefined [Latin American students studying in Germany]). There were 7 samples from Australia and 1 from Oceania (excluding Australia): Fiji Island. Only 6 samples were Cross-Cultural (containing samples from more than five different countries). Combining the figures, we found that around 8 out of 10 samples were from WEIRD populations (81%, Europe/North America/Australia), and that 87% of the samples used were from developed regions (following the UN classification; United Nations 2013).

In terms of sample type, 113 of the 311 samples were Western student samples, while 24 were non-Western student samples; 60 samples were online paid crowdsourced, while 20 were unpaid crowdsourced. Thus, 70% of the samples were either online samples or student samples. Twenty-five samples were based on children (21 from Western and 4 from non-Western populations). Only a small fraction of the samples consisted of non-Western adults who were not students (24 out of 311 samples, or 8%).

Fig. 1 Origin of samples. N. America, North America; L. America, Latin America and Caribbean

Are Samples from Certain Geographical Locations Larger than Others?

Given that there was only one sample from Oceania (excluding Australia) (see Figs. 1 and 2), we combined this with Australia for the analysis of variation in sample sizes between regions (see ESM for additional analyses using this combination). Variation in sample size between geographical regions was not statistically significant (Kruskal-Wallis test: $\chi^2(6) = 10.095$, $p = .12$). Following adjustment for multiple testing, the median sample size was found to be significantly larger for cross-cultural samples (which, according to our coding criteria, had to contain data from more than five different countries) than for Latin American and Caribbean samples ($p = .037$). The ESM contains all post-hoc multiple comparisons (all remaining p values > .09; see ESM).

Are some Types of Samples Larger than Others?

The sample sizes differed significantly according to type (Fig. 3; Kruskal-Wallis test: $\chi^2(7) = 63.9$, $p < .0001$). Post-hoc comparisons adjusted for multiple testing using the Benjamini-Hochberg procedure showed that online (paid crowdsourced) and online (unpaid crowdsourced) samples tended to be larger than other types of samples (Fig. 3; Table 1).

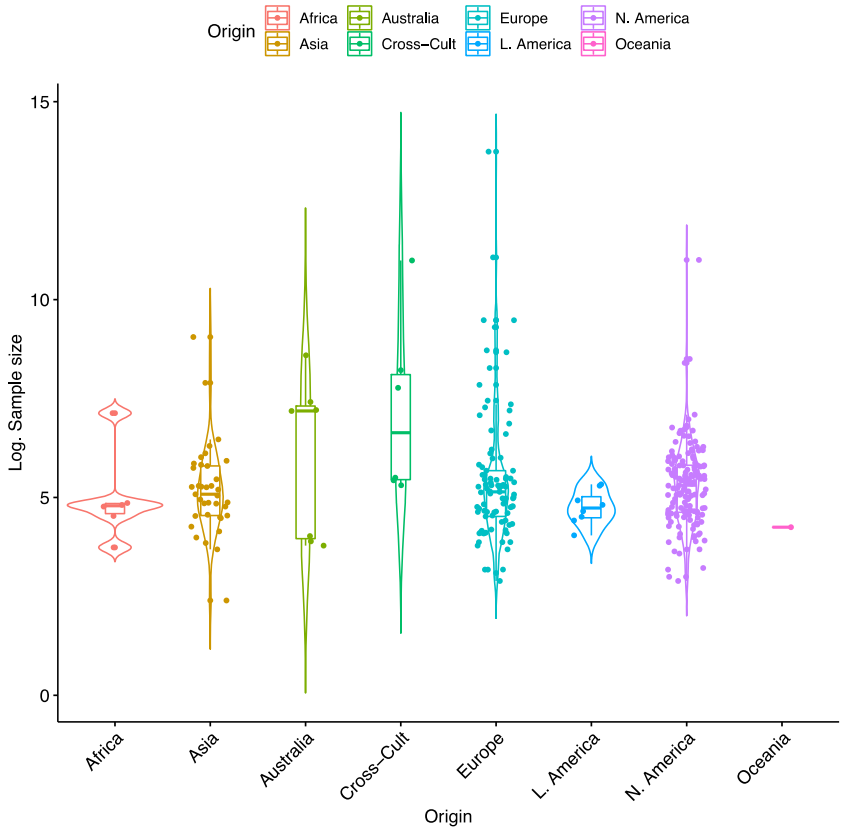

Fig. 2 Violin plot for geographical origin and Log. sample size, density distribution (curve), median (horizontal line), interquartile range (IQR, box), whiskers (1.5 times the IQR) and individual samples (dots)

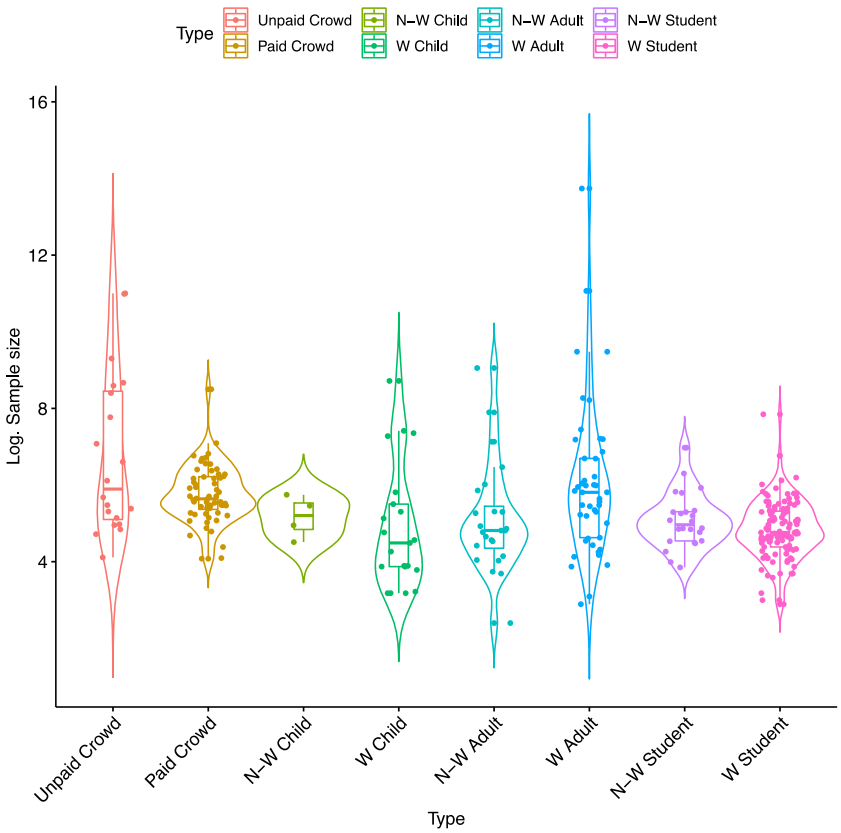

Fig. 3 Violin plot for geographical origin and Log. sample size, density distribution (curve), median (horizontal line), interquartile range (IQR, box), whiskers (1.5 times the IQR) and individual samples (dots)

Discussion

Our survey of papers published in 2015–2016 in two key journals relevant to evolutionary psychology, '*Evolution & Human Behavior*' and '*Evolutionary Psychology*', indicated a clear dominance of adult samples from Western, developed countries, with a particular preponderance of North American samples. Seventy percent of samples were sourced online or from student populations. Asian samples mainly consisted of samples from Japan. Notable under-representations included samples collected from Africa and Latin America (including the Caribbean). Data collected online typically gave rise to the largest sample sizes.

Implications of Relying on WEIRD Samples

In our survey, 81% of samples were from WEIRD populations. The main advantages of relying on WEIRD samples arise from the fact that most authors are WEIRD, and so WEIRD samples are more practical and convenient, particularly in respect of ease of access and low costs of data sampling. Requiring costly and time-consuming data collection can stifle scientific research, given that it is often poorly resourced (Lakens et al. 2018). There are obvious practical difficulties in collecting data outside of one's country of residence, or in countries where attitudes and familiarity may vary in relation to the psychological procedures that will be well-known to readers of this journal, such as models of obtaining informed consent, or methods of data elicitation. Indeed, we do not always need to go far to find diverse samples (e.g. Hill et al. 2014; Nettle 2017; Wilson 2011; Wilson et al. 2009). WEIRD samples themselves are certainly not homogeneous, and for instance even people from different neighbourhoods within the same city can vary as much as people from entirely different cultures (Nettle 2017; Nettle et al. 2011). Thus, even samples from WEIRD cultures can be sufficiently diverse that they provide some useful evidence in support of generalisability.

On the other hand, evolutionary psychologists are often keen to sample outside WEIRD populations because of their interest in assessing and recording the nature of humans as a species. WEIRD populations may experience environments (in terms of novel technology, experience of hunger, exposure to death and so on) that are particularly dissimilar from those experienced by many of our ancestors, something that needs to be borne in mind when constructing and testing evolutionary theories of behaviour. Human universals (Brown 1991, 2000) may only be uncovered following assessment of multiple human populations, if not all populations, and assist in developing and evaluating evolutionary theories of behaviour. If we are to

Table 1 Post-hoc comparisons based on the type of sample

	Group 1	Group 2	Adjusted *p* value
1	W student	Online paid	< .00001
2	W student	Online unpaid	0.00018
3	W student	W adult	0.00018
4	N-W student	Online paid	0.00096
5	N-W adult	Online paid	0.00518
6	W child	Online paid	0.00582
7	W child	Online unpaid	0.00797
8	N-W adult	Online unpaid	0.00826
9	N-W student	Online unpaid	0.00858
10	W adult	W child	0.02981
11	N-W student	W adult	0.04100
12	W adult	N-W adult	0.06205
13	N-W child	Online paid	0.22546
14	N-W child	Online unpaid	0.34080
15	N-W student	W child	0.34080
16	W student	N-W student	0.41799
17	Online paid	Online unpaid	0.43369
18	W adult	Online unpaid	0.43369
19	W child	N-W child	0.43369
20	W adult	N-W child	0.43369
21	W student	N-W child	0.43369
22	N-W adult	W child	0.43369
23	W student	W child	0.43369
24	N-W adult	N-W child	0.62156
25	N-W student	N-W adult	0.70330
26	N-W student	N-W child	0.72102
27	W student	N-W adult	0.73921
28	W adult	Online paid	0.94837

test how individual differences may be functionally adaptive (Tybur et al. 2014; Wilson 1998), we need to compare individuals from different ecological settings (Nettle 2009). An awareness of the diversity of worldwide human behaviours would help us understand how human behaviours emerge from an interaction between local ecologies and our evolved brains (Henrich et al. 2010). For these sorts of reasons, classic studies that seek to test adaptive reasoning have taken pains to survey different populations (e.g. Buss 1989; Daly and Wilson 1988; Kenrick and Keefe 1992; Schaller and Murray 2008; Schmitt 2005; Scott et al. 2014). Reliance on WEIRD populations limits discovery of any patterns that might allow us to predict domains where psychological phenomena are more likely to be universal, and domains where psychological phenomena are more likely to show variability (Henrich et al. 2010). As an additional step, WEIRD authors (including ourselves) could usefully reach out to non-WEIRD collaborators to attempt to draw from wider samples. Encouraging greater diversity among authors should automatically increase participant diversity (Medin et al. 2017).

Henrich et al.'s (2010) renowned position piece explains that participants from WEIRD (Western, Educated, Industrialised, Rich, Democratic) populations can be more or less universally representative, dependent upon the area of research, and goes on to detail where reliance on WEIRD populations might not present a complete picture. To summarise Henrich et al.'s findings in as far as they are of particular concern to evolutionary psychologists, behavioural economics games used to assess fairness and co-operation showed that western undergraduate samples behaved very differently from participants from other societies. Similarly, folkbiological reasoning develops differently in rural American children compared with children from other settings. Further, research on moral reasoning has also now shown significant differences between the original data collected from western cultures and data collected later among more diverse cultures. In each case, theories were initially developed that assumed that the results from cultures more familiar to the researchers were universal. On the other hand, some research topics that will be familiar to readers of this journal seem to be those for which we have good evidence of universality, and where a reliance on less diverse samples is less problematic. Such topics can include emotional expression and pride displays, false belief tasks, some mate preferences, personality structure, psychological essentialism, punishment of free-riders and social relationships (Henrich et al. 2010).

Implications of Relying on Student Samples

We found that 44% of the samples that we coded were student samples (137 out of 311 samples). The advantages of using student samples are similar to the advantages of using WEIRD samples, together with the additional advantages that student participants should be comfortable within the university setting and accustomed to following task instructions (Rosenthal 1965). On the other hand, a reliance on student samples may be particularly problematic when dealing with topics where there is a clear impact of the variables that distinguish students from the general population. These may include broad variables such as age, experience, socio-economic background and educational level, as well as more specific tendencies including students' greater level of cognitive ability and obedience to authority, more transient friendships and still nascent attitudes and sense of self (Sears 1986). Areas that have received specific criticism due to their reliance on student sampling include research on economic decision making (Levitt and List 2007), socio-political attitudes (Schultz 1969), the psychological processes relating to prejudice (Henry 2008) and industrial and organisational psychology (Bergman and

Jean 2016). Effect sizes calculated from student data can differ from other populations not merely in magnitude, but also in direction (Peterson 2001). Further, if researchers are interested in features of a variable (e.g. its range, distribution, mean), then it will not be possible to assess that accurately from a sample that is partially selected in relation to that variable: thus, population-level IQ scores cannot be assessed from student samples.

Implications of Relying on Online Samples

Our survey pointed to a substantial reliance (around one-quarter of all samples) on online data collection. It has been suggested that the internet offers a practical solution to reliance upon WEIRD samples (Gosling et al. 2004). Advantages of online sampling include cheap, quick and convenient access to participants who can often be recruited in larger numbers than are readily available for offline studies, and this ease of access to large online samples appears to be reflected in our analyses above. Indeed, online sampling can reach a greater diversity of participants, including some difficult-to-reach and under-represented populations (Andrews et al. 2003). Further, the anonymous setting of an online survey might arguably provoke more honest answers to questions on sensitive topics, compared with lab data collection (Joinson 1999). From another perspective though, internet sampling is limited in terms of the kinds of research tools that can be used, and in addition internet access itself is only available to a proportion of the population (and in some countries, a smaller proportion of the population than those who constitute undergraduate samples in other countries), meaning that the sample is still restricted (Gosling et al. 2010). Researchers sometimes raise concerns that data collected via online sampling might be of lower quality than that collected using more traditional methods (e.g. Matzat and Snijders 2010; Paolacci and Chandler 2014). Online participants do not have easy access to the researcher to raise queries, might enter data carelessly or thoughtlessly, or might have chosen to enter data merely in order to view the study content (Aust et al. 2013). Accordingly, to test the quality of online data collection, various studies have compared data collected online and offline and concluded that in many instances the two sampling methodologies give rise to very similar outcomes (Krantz and Dalal 2000). For instance, judgements of the attractiveness of different female body shapes were similar, irrespective of whether data were collected from laboratory studies of psychology undergraduate students or online from visitors to psychology webpages hosted by the same university (Krantz et al. 1997). Nevertheless, data collected online and offline are not identical (Birnbaum 2004; Epstein et al. 2001). This is unsurprising, given that context and environment can influence behaviour. The demographic differences between people with and without internet access may be particularly stark in developing countries, and so online sampling may not be the most suitable way to reach diverse populations in those countries (Batres and Perrett 2014), and in some instances of course internet access can contribute to behaviour that we might want to assess; for example, media exposure appears to explain differences in preferences for faces and body types (Boothroyd et al. 2016). However, despite the differences between online and offline samples, one should not be seen as the poor cousin of the other; both have strengths and can be used in complementary fashion to rigorously test inferences.

Implications of Relying on Restricted Samples

From a statistical point of view, restricted sampling can lead to selection bias which in turn can lead to confounding (Bareinboim and Pearl 2012; Elwert and Winship 2014; Fiedler 2000; Rohrer 2018). Recently, statisticians have more explicitly defined the conditions under which causal inferences can be made when combining data under heterogeneous conditions (Bareinboim and Pearl 2016). Depending on the type of inference researchers want to make, they could face confounding, sampling bias, or transportability bias. Importantly, these issues apply to both the decision to focus on (for example) a WEIRD population as well as expanding the research to non-WEIRD populations. An exclusive focus on restricted samples comes at the cost of external validity. In medical research, there has been a repeated call to revalue external validity (e.g. Burchett et al. 2011; Green and Glasgow 2006; Steckler and McLeroy 2008). While there had previously been a strong focus on internal validity, for example, focussing on questions such as whether confounding can be effectively ruled out in randomised controlled trials, there has been a call to also remember the importance of external validity (can we generalise the findings from this trial?). An obvious issue with relying on WEIRD, student and/or online samples would be the degree to which any conclusions would hold in different populations (Henrich et al. 2010; Henry 2008; Sue 1999).

Even more fundamentally, and before making causal inferences about other populations, researchers face the more basal problem of knowing whether they are measuring the same 'thing' in different populations. This issue is well-understood in the field of psychometrics and has led to the development of measurement techniques and tests to examine the degree to which constructs are measured consistently across cultures (Heine et al. 2002; Hui and Triandis 1985; Nasif et al. 1991; Poortinga 1989; Van de Vijver and Leung 1997). We did not explicitly assess how many papers established equivalence of measurement between different samples, as we focused on the samples. Our standard psychological instruments, often developed by researchers working

within WEIRD settings, may limit the generalisability of research findings (Ceci et al. 2010; Konečni 2010; Rochat 2010). We therefore call for more research explicitly establishing that the same 'thing' is measured in different populations. Depending on the sampling scheme, broadening the research to non-WEIRD populations could also give rise to problems such as non-independence (Mace and Pagel 1994; Naroll 1965; Pollet et al. 2014; Ross and Homer 1976), which then would need to be addressed. We do not discuss these issues in further detail here, as the degree to which they matter could differ on design (experiment/correlational), covariates and research question. For example, for many psychophysical studies, and also evolutionary psychological studies (Tybur et al. 2014), the focus is on within-individual differences. The implicit assumption is that these would not vary depending on the population studied. For such studies, it would be useful for authors to be more explicit to which degree these within-individual differences are expected to generalise to other samples. In some cases, restricted sampling in itself is useful to determine whether a behaviour exists or not, and as such testing a WEIRD, student and/or online population could constitute a necessary first step (Greenwood 1982; Mook 1983). Given that every sample is restricted in some way, authors can usefully make a statement pertaining to the constraints on generality, to explain the boundaries of the population that they believe their results to apply to (Simons et al. 2017). More broadly, the field would benefit from setting out the conditions under which causal inferences can be made (Pearl 2009a, b).

Limitations and Future Directions

Our analysis did not cover all of the journals that publish evolutionary psychological studies. Instead, mirroring the work that has been done in other reviews of sample diversity (e.g. Arnett 2008; Gallander Wintre et al. 2001; Sears 1986), we focussed on key journals. Many papers on evolutionary psychology are published outside of those two journals. Similarly, there might be papers in our sample which have a different focus than evolutionary psychology, or which might be better classified as relating to fields such as comparative cognition, behavioural economics, linguistics, demography, or anthropology among others. Future work might compare one journal to the next, and compare across a sequence of different years, to determine the variance in sample diversity. Alternatively, one could define keywords to more clearly delineate articles covering evolutionary psychology. Further, future research might seek to uncover whether different research areas within evolutionary psychology are more or less reliant upon non-diverse samples, and how this corresponds to their development as a research area. Exploratory studies might well choose to focus on easily accessible samples such as undergraduates to test their initial ideas, whereas more mature research areas ought to seek to diversify their samples further in order to test the generalisability of their findings. Even within specific research areas, some topics have been studied in more diverse worldwide samples than others; for instance, sex differences in partner preferences draws from data from many cultures (e.g. Buss 1989; Shackelford et al. 2005), whereas research on ovulatory shifts in partner preferences very much rests upon studies carried out in English-speaking WEIRD countries (Gildersleeve et al. 2014).

Our survey does not present cause for despair. In terms of participant diversity, evolutionary psychology does rely on WEIRD, student samples less heavily than some fields (Apicella and Barrett 2016; Kurzban 2013). However, this is perhaps in part because of the discipline's need for cross-cultural surveys to validate theories that claim to purport to humans as a species. Evolutionary psychology has a greater need for cross-cultural replications than other disciplines, such as those focussed around basic psychophysics where we might more easily assume universal underlying mechanics, or more descriptive research approaches that aim to uncover behaviour in culturally-specific environments, such as the workplace or social media sites. We do not mean to imply either that sample diversity should be the only goal; there are many valuable ways to add to our understanding of any phenomenon. Valuable extensions to research on a WEIRD, student sample can arise, for instance, from adding methodological diversity, developing theoretical frameworks, creating models of the behaviour, or testing similar behaviours in other species. Developmental approaches can make a useful testing ground for adaptive predictions, given that individuals have different adaptive needs across their life course, but we note that only 8% of the samples that we coded used child participants. Scientists, including evolutionary psychologists, are increasingly recognising the value in replication across multiple labs and samples (e.g. Camerer et al. 2016; Ebersole et al. 2016; Errington et al. 2014; Zwaan et al. 2018). In this light, it is of interest that the first study to be accepted by the Psychological Science Accelerator (https://psysciacc.wordpress.com/), a project that uses multiple laboratories to test hypotheses, was proposed by two researchers, Jones and DeBruine, whose work often draws upon a functional framework. For now, we conclude that while two key journals use more diverse samples than many typical (social or developmental) psychology journals, as Kurzban (2013) suggested, it is important to realise that given the glaring

underrepresentation of certain regions we still have a long road ahead.

Acknowledgements Part of the data were collected while the first author was at the University of Leiden, and he wishes to thank his bachelor thesis group for their support with the project. We thank the editor and two reviewers for their very helpful input.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

Adair, J. G., Coêlho, A. E. L., & Luna, J. R. (2002). How international is psychology? *International Journal of Psychology, 37*(3), 160–170. https://doi.org/10.1080/00207590143000351.

Andrews, D., Nonnecke, B., & Preece, J. (2003). Electronic survey methodology: a case study in reaching hard-to-involve internet users. *International Journal of Human-Computer Interaction, 16*(2), 185–210. https://doi.org/10.1207/S15327590IJHC1602_04.

Apicella, C. L., & Barrett, H. C. (2016). Cross-cultural evolutionary psychology. *Current Opinion in Psychology, 7*, 92–97. https://doi.org/10.1016/j.copsyc.2015.08.015.

Arnett, J. J. (2008). The neglected 95%: why American psychology needs to become less American. *American Psychologist, 63*(7), 602–614. https://doi.org/10.1037/0003-066X.63.7.602.

Aust, F., & Barth, M. (2016). papaja: Create *APA manuscripts with R Markdown*. See https://github.com/crsh/papaja.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2.

Bareinboim, E., & Pearl, J. (2012). Controlling selection bias in causal inference. In *Proceedings of the fifteenth international conference on artificial intelligence and statistics* (pp. 100–108).

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences, 113*(27), 7345 LP–7347352. https://doi.org/10.1073/pnas.1510507113.

Batres, C., & Perrett, D. I. (2014). The influence of the digital divide on face preferences in El Salvador: people without internet access prefer more feminine men, more masculine women, and women with higher adiposity. *PLoS One, 9*(7), e100966. https://doi.org/10.1371/journal.pone.0100966.

Bauserman, R. (1997). International representation in the psychological literature. *International Journal of Psychology, 32*(2), 107–112. https://doi.org/10.1080/002075997400908.

Bello, D., Leung, K., Radebaugh, L., Tung, R. L., & van Witteloostuijn, A. (2009). From the editors: student samples in international business research. *Journal of International Business Studies, 40*(3), 361–364. https://doi.org/10.1057/jibs.2008.101.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300. https://doi.org/10.2307/2346101.

Bergman, M. E., & Jean, V. A. (2016). Where have all the "workers" gone? A critical analysis of the unrepresentativeness of our samples relative to the labor market in the industrial–organizational psychology literature. *Industrial and Organizational Psychology, 9*(1), 84–113. https://doi.org/10.1017/iop.2015.70.

Berry, J. W. (2013). Achieving a global psychology. *Canadian Psychology/Psychologie Canadienne, 54*(1), 55–61. https://doi.org/10.1037/a0031246.

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology, 55*(1), 803–832. https://doi.org/10.1146/annurev.psych.55.090902.141601.

Bohannon, J. (2016). Psychologists grow increasingly dependent on online research subjects. *Science*. https://doi.org/10.1126/science.aag0592.

Boothroyd, L. G., Jucker, J., Thornborrow, T., Jamieson, M. A., Burt, D. M., Barton, R. A., et al. (2016). Television exposure predicts body size ideals in rural Nicaragua. *British Journal of Psychology, 107*(4), 752–767. https://doi.org/10.1111/bjop.12184.

Brown, D. E. (1991). *Human universals*. New York, NY: McGraw-Hill.

Brown, D. E. (2000). Human universals and their implications. In N. Roughley (Ed.), *Being humans: anthropological universality and particularity in transdisciplinary perspectives* (pp. 156–174). Berlin: Walter de Gruyter.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. https://doi.org/10.1177/1745691610393980.

Burchett, H., Umoquit, M., & Dobrow, M. (2011). How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. *Journal of Health Services Research & Policy, 16*(4), 238–244. https://doi.org/10.1258/jhsrp.2011.010124.

Buss, D. M. (1989). Sex differences in human mate preferences: evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences, 12*(1), 1–49. https://doi.org/10.1017/S0140525X00023992.

Buss, D. M. (1994). *The evolution of desire: Strategies of human mating*. New York: Basic books.

Buss, D. M. (1995). Evolutionary psychology: a new paradigm for psychological science. *Psychological Inquiry, 6*(1), 1–30. https://doi.org/10.1207/s15327965pli0601_1.

Buss, D. M., & Shackelford, T. K. (1997). From vigilance to violence: mate retention tactics in married couples. *Journal of Personality and Social Psychology, 72*(2), 346–361. https://doi.org/10.1037/0022-3514.72.2.346.

Buss, D. M., Larsen, R. J., Westen, D., & Semmelroth, J. (1992). Sex differences in jealousy: evolution, physiology, and psychology. *Psychological Science, 3*(4), 251–255. https://doi.org/10.1111/j.1467-9280.1992.tb00038.x.

Buss, D. M., Shackelford, T. K., & LeBlanc, G. J. (2000). Number of children desired and preferred spousal age difference: context-specific mate preference patterns across 37 cultures. *Evolution and Human Behavior, 21*(5), 323–331. https://doi.org/10.1016/S1090-5138(00)00048-9.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918.

Ceci, S. J., Kahan, D. M., & Braman, D. (2010). The WEIRD are even weirder than you think: diversifying contexts is as important as

diversifying samples. *Behavioral and Brain Sciences, 33*(2–3), 87–88. https://doi.org/10.1017/S0140525X10000063.

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics, 2*(4), 205316801562207. https://doi.org/10.1177/2053168015622072.

Cole, M. (2006). Internationalism in psychology: we need it now more than ever. *American Psychologist, 61*(8), 904–917. https://doi.org/10.1037/0003-066X.61.8.904.

Cosmides, L., & Tooby, J. (1997). *Evolutionary psychology: A primer*. http://www.cep.ucsb.edu/primer.html.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410.

Daly, M., & Wilson, M. (1988). *Homicide*. New Brunswick: Transaction Books.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., Davis, W. E., Devos, T., Fletcher, M. M., German, K., Grahe, J. E., Hermann, A. D., Hicks, J. A., Honeycutt, N., Humphrey, B., Janus, M., Johnson, D. J., Joy-Gaba, J. A., Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, R. A., Lucas, R. E., Lustgraaf, C. J. N., Martin, D., Menon, M., Metzger, M., Moloney, J. M., Morse, P. J., Prislin, R., Razza, T., Re, D. E., Rule, N. O., Sacco, D. F., Sauerberger, K., Shrider, E., Shultz, M., Siemsen, C., Sobocko, K., Weylin Sternglanz, R., Summerville, A., Tskhay, K. O., van Allen, Z., Vaughn, L. A., Walker, R. J., Weinberg, A., Wilson, J. P., Wirth, J. H., Wortman, J., & Nosek, B. A. (2016). Many labs 3: evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: the problem of conditioning on a collider variable. *Annual Review of Sociology, 40*, 31–53. https://doi.org/10.1146/annurev-soc-071913-043455.

Epstein, J., Klinkenberg, W., Wiley, D., & McKinley, L. (2001). Insuring sample equivalence across internet and paper-and-pencil assessments. *Computers in Human Behavior, 17*(3), 339–346. https://doi.org/10.1016/S0747-5632(01)00002-4.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *eLife, 3*. https://doi.org/10.7554/eLife.04333.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*(4), 659–676. https://doi.org/10.1037/0033-295X.107.4.659.

Gallander Wintre, M., North, C., & Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: a developmental perspective. *Canadian Psychology/Psychologie Canadienne, 42*(3), 216–225. https://doi.org/10.1037/h0086893.

Gildersleeve, K., Haselton, M. G., & Fales, M. R. (2014). Do women's mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychological Bulletin, 140*(5), 1205–1259. https://doi.org/10.1037/a0035438.

Gosling, S. D., & Johnson, J. A. (2010). In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research*. Washington, DC: American Psychological Association. https://doi.org/10.1037/12076-000.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*(2), 93–104. https://doi.org/10.1037/0003-066X.59.2.93.

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: the promise of the internet in reaching more diverse samples. *Behavioral and Brain Sciences, 33*(2–3), 94–95. https://doi.org/10.1017/S0140525X10000300.

Graham, S. (1992). "Most of the subjects were White and middle class": trends in published research on African Americans in selected APA journals, 1970-1989. *American Psychologist, 47*(5), 629–639. https://doi.org/10.1037/0003-066X.47.5.629.

Green, L. W., & Glasgow, R. E. (2006). Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. *Evaluation & the Health Professions, 29*(1), 126–153. https://doi.org/10.1177/0163278705284445.

Greenwood, J. D. (1982). On the relation between laboratory experiments and social behaviour: causal explanation and generalization. *Journal for the Theory of Social Behaviour, 12*(3), 225–250. https://doi.org/10.1111/j.1468-5914.1982.tb00449.x.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: the reference-group effect. *Journal of Personality and Social Psychology, 82*(6), 903–918. https://doi.org/10.1037//0022-3514.82.6.903.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28*(06), 795–815. https://doi.org/10.1017/S0140525X05000142.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X.

Henry, P. J. (2008). College sophomores in the laboratory redux: influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry, 19*(2), 49–71. https://doi.org/10.1080/10478400802049936.

Hill, J. M., Jobling, R., Pollet, T. V., & Nettle, D. (2014). Social capital across urban neighborhoods: a comparison of self-report and observational data. *Evolutionary Behavioral Sciences, 8*(2), 59–69. https://doi.org/10.1037/h0099131.

Hlavac, M. (2014). *Stargazer: LaTeX code and ASCII text for well-formatted regression and summary statistics tables*. http://cran.r-project.org/web/packages/stargazer/index.html.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology. *Journal of Cross-Cultural Psychology, 16*(2), 131–152. https://doi.org/10.1177/0022002185016002001.

Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers, 31*(3), 433–438. https://doi.org/10.3758/BF03200723.

Keene, O. N. (1995). The log transformation is special. *Statistics in Medicine, 14*(8), 811–819. https://doi.org/10.1002/sim.4780140810.

Kenrick, D. T., & Keefe, R. C. (1992). Age preferences in mates reflect sex differences in human reproductive strategies. *Behavioral and Brain Sciences, 15*(01), 75–91. https://doi.org/10.1017/S0140525X00067595.

Kenrick, D. T., Keefe, R. C., Bryan, A., Barr, A., & Brown, S. (1995). Age preferences and mate choice among homosexuals and heterosexuals: a case for modular psychological mechanisms. *Journal of Personality and Social Psychology, 69*(6), 1166–1172. https://doi.org/10.1037/0022-3514.69.6.1166.

Konečni, V. J. (2010). Responsible behavioral science generalizations and applications require much more than non-WEIRD samples. *Behavioral and Brain Sciences, 33*(2–3), 98–99. https://doi.org/10.1017/S0140525X10000142.

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In *Psychological experiments on the internet* (pp. 35–60).

Amsterdam: Elsevier. https://doi.org/10.1016/B978-012099980-4/50003-4.

Krantz, J. H., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, & Computers, 29*(2), 264–269. https://doi.org/10.3758/BF03204824.

Kurzban, R. (2013). Is evolutionary psychology weird or normal? *EP Journal blog*. http://epjournal.net/blog/2013/09/is-evolutionary-psychology-weird-or-normal/. Accessed 14 Nov 2016.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., van Calster, B., Carlsson, R., Chen, S. C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggenberger, R., Grist, J., van Harmelen, A. L., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A., Konat, B., Lao, J., Larsen, E. G., Lodder, G. M. A., Lukavský, J., Madan, C. R., Manheim, D., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q. X., Nilsonne, G., de Oliveira, C. L., de Xivry, J. J. O., Parsons, S., Pfuhl, G., Quinn, K. A., Sakon, J. J., Saribay, S. A., Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smits, T., Spies, J. R., Sreekumar, V., Steltenpohl, C. N., Stenhouse, N., Świątkowski, W., Vadillo, M. A., van Assen, M. A. L. M., Williams, M. N., Williams, S. E., Williams, D. R., Yarkoni, T., Ziano, I., & Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour, 2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21*(2), 153–174. https://doi.org/10.1257/jep.21.2.153.

Mace, R., & Pagel, M. (1994). The comparative method in anthropology. *Current Anthropology, 35*(5), 549–564. https://doi.org/10.1086/204317.

Matzat, U., & Snijders, C. (2010). Does the online collection of ego-centered network data reduce data quality? An experimental comparison. *Social Networks, 32*(2), 105–111. https://doi.org/10.1016/j.socnet.2009.08.002.

Medin, D., Ojalehto, B., Marin, A., & Bang, M. (2017). Systems of (non-)diversity. *Nature Human Behaviour, 1*, 88. https://doi.org/10.1038/s41562-017-0088.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379–387. https://doi.org/10.1037/0003-066X.38.4.379.

Müller, K. (2017). *bindrcpp*. https://cran.r-project.org/web/packages/bindrcpp/index.html.

Naroll, R. (1965). Galton's problem: the logic of cross-cultural analysis. *Social Research, 32*(4), 428–451.

Nasif, E. G., Al-Daeaj, H., Ebrahimi, B., & Thibodeaux, M. S. (1991). Methodological problems in cross-cultural research: an updated review. *MIR: Management International Review, 31*(1), 79–91.

Nettle, D. (2009). Ecological influences on human behavioural diversity: a review of recent findings. *Trends in Ecology & Evolution, 24*(11), 618–624. https://doi.org/10.1016/j.tree.2009.05.013.

Nettle, D. (2017). *Tyneside neighbourhoods: Deprivation, social life and social behaviour in one British city*. Cambridge: Open Book Publishers.

Nettle, D., Colléony, A., & Cockerill, M. (2011). Variation in cooperative behaviour within a single city. *PLoS One, 6*(10), e26922. https://doi.org/10.1371/journal.pone.0026922.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: a call to action. *Journal of Experimental Child Psychology, 162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017.

Paolacci, G., & Chandler, J. (2014). Inside the turk understanding mechanical turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184–188. https://doi.org/10.1177/0963721414531598.

Pearl, J. (2009a). *Causality*. Cambridge: Cambridge University Press.

Pearl, J. (2009b). Causal inference in statistics: an overview. *Statistics Surveys, 3*, 96–146. https://doi.org/10.1214/09-SS057.

Peterson, R. A. (2001). On the use of college students in social science research: insights from a second-order meta-analysis. *Journal of Consumer Research, 28*(3), 450–461. https://doi.org/10.1086/323732.

Pollet, T. V., Tybur, J. M., Frankenhuis, W. E., & Rickard, I. J. (2014). What can cross-cultural correlations teach us about human nature? *Human nature (Hawthorne, N.Y.), 25*(3), 410–429. https://doi.org/10.1007/s12110-014-9206-3.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: an overview of basic issues. *International Journal of Psychology, 24*(1), 737–756. https://doi.org/10.1080/00207598908247842.

R Development Core Team. (2008). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raihani, N. J., Mace, R., & Lamba, S. (2013). The effect of $1, $5 and $10 stakes in an online dictator game. *PLoS One, 8*(8), e0073131. https://doi.org/10.1371/journal.pone.0073131.

Robinson, D. (2017). *Broom*. https://cran.r-project.org/web/packages/broom/index.html.

Rochat, P. (2010). What is really wrong with a priori claims of universality? Sampling, validity, process level, and the irresistible drive to reduce. *Behavioral and Brain Sciences, 33*(2–3), 107–108. https://doi.org/10.1017/S0140525X10000233.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science, 1*(1), 27–42. https://doi.org/10.1177/2515245917745629.

Rosenthal, R. (1965). The volunteer subject. *Human Relations, 18*(4), 389–406.

Ross, M. H., & Homer, E. (1976). Galton's problem in cross-national research. *World Politics, 29*(1), 1–28. https://doi.org/10.2307/2010045.

Schaller, M., & Murray, D. R. (2008). Pathogens, personality, and culture: disease prevalence predicts worldwide variability in sociosexuality, extraversion, and openness to experience. *Journal of Personality and Social Psychology, 95*(1), 212–221. https://doi.org/10.1037/0022-3514.95.1.212.

Schmitt, D. P. (2005). Sociosexuality from Argentina to Zimbabwe: a 48-nation study of sex, culture, and strategies of human mating. *Behavioral and Brain Sciences, 28*(2), 247–275. https://doi.org/10.1017/S0140525X0500005.

Schultz, D. P. (1969). The human subject in psychological research. *Psychological Bulletin, 72*(3), 214–228. https://doi.org/10.1037/h0027880.

Scott, I. M., Clark, A. P., Josephson, S. C., Boyette, A. H., Cuthill, I. C., Fried, R. L., Gibson, M. A., Hewlett, B. S., Jamieson, M., Jankowiak, W., Honey, P. L., Huang, Z., Liebert, M. A., Purzycki, B. G., Shaver, J. H., Snodgrass, J. J., Sosis, R., Sugiyama, L. S., Swami, V., Yu, D. W., Zhao, Y., & Penton-Voak, I. S. (2014). Human preferences for sexually dimorphic faces may be evolutionarily novel. *Proceedings of the National Academy of Sciences, 111*(40), 14388–14393. https://doi.org/10.1073/pnas.1409643111.

Scott-Phillips, T. C. (2018). *Expression unleashed*. https://osf.io/umk9g/.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*(3), 515–530. https://doi.org/10.1037/0022-3514.51.3.515

Shackelford, T. K., Schmitt, D. P., & Buss, D. M. (2005). Universal dimensions of human mate preferences. *Personality and Individual Differences, 39*(2), 447–458. https://doi.org/10.1016/j.paid.2005.01.023.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences. (2nd ed.).* New York, NY: McGraw-hill.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): a proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630.

Smart, R. G. (1966). Subject selection bias in psychological research. *Canadian Psychologist/Psychologie canadienne, 7a*(2), 115–121. https://doi.org/10.1037/h0083096.

Steckler, A., & McLeroy, K. R. (2008). The importance of external validity. *American Journal of Public Health, 98*(1), 9–10. https://doi.org/10.2105/AJPH.2007.126847.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651.

Stulp, G., Simons, M. J. P., Grasman, S., & Pollet, T. V. (2017). Assortative mating for human height: a meta-analysis. *American Journal of Human Biology, 29*(1), e22917. https://doi.org/10.1002/ajhb.22917.

Sue, S. (1999). Science, ethnicity, and bias: where have we gone wrong? *American Psychologist, 54*(12), 1070–1077. https://doi.org/10.1037/0003-066X.54.12.1070.

Tooby, J., & Cosmides, L. (1990). On the universality of human nature and the uniqueness of the individual: the role of genetics and adaptation. *Journal of Personality, 58*(1), 17–67. https://doi.org/10.1111/j.1467-6494.1990.tb00907.x.

Tybur, J. M., Frankenhuis, W. E., & Pollet, T. V. (2014). Behavioral immune system methods: surveying the present to shape the future. *Evolutionary Behavioral Sciences, 8*(4), 274–283. https://doi.org/10.1037/ebs0000017.

United Nations (2013). *Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings*. https://unstats.un.org/unsd/methods/m49/m49regin.htm.

Usher, E. L. (2018). Acknowledging the whiteness of motivation research: seeking cultural relevance. *Educational Psychologist, 53*(2), 131–144. https://doi.org/10.1080/00461520.2018.1442220.

Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed.). Boston: Allyn & Bacon.

Wickham, H. (2009). *ggplot2*. New York: Springer New York. https://doi.org/10.1007/978-0-387-98141-3.

Wickham, H. (2014). *Tidyr: Easily tidy data with spread and gather functions*. http://cran.r-project.org/web/packages/tidyr.

Wickham, H., & Bryan, J. (2017). *readxl: read excel files*. R package version 1.0.0. https://cran.r-project.org/package=readxl.

Wickham, H., & Francois, R. (2017). *dplyr: a grammar of data manipulation*. http://dplyr.tidyverse.org/.

Wickham, H., & Wickham, M. H. (2017). *'plyr' package*. https://cran.r-project.org/package=plyr.

Wilson, D. S. (1998). Adaptive individual differences within single populations. *Philosophical Transactions of the Royal Society B: Biological Sciences, 353*(1366), 199–205. https://doi.org/10.1098/rstb.1998.0202.

Wilson, D. S. (2011). *The neighborhood project: Using evolution to improve my city, one block at a time*. New York: Little, Brown.

Wilson, D. S., O'Brien, D. T., & Sesma, A. (2009). Human prosociality from an evolutionary perspective: variation and correlations at a city-wide scale. *Evolution and Human Behavior, 30*(3), 190–200. https://doi.org/10.1016/j.evolhumbehav.2008.12.002.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, FL: CRC Press.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, e120. https://doi.org/10.1017/S0140525X17001972.