No Evidence Against the Greater Male Variability Hypothesis: A Commentary on Harrison et al.'s (2022) Meta-Analysis of Animal Personality

Marco Del Giudice

Steven W. Gangestad

University of New Mexico

In press (2023). ***Evolutionary Psychological Science*****.**

Address correspondence to Marco Del Giudice, Department of Psychology, University of New Mexico. Logan Hall, 2001 Redondo Dr. NE, Albuquerque, NM 87131, USA; email: marcodg@unm.edu

Abstract

In a recent meta-analysis, Harrison, Noble and Jennions (2022; *Biological Reviews, 97,* 679-707) set out to test the greater male variability hypothesis with respect to personality in non-human animals. Based on their non-significant results, they concluded that there is no evidence to support the hypothesis, and that biological explanations for greater male variability in human psychological traits should be called into question. Here, we show that these conclusions are unwarranted. Specifically: (a) in mammals, birds, and reptiles/amphibians, the magnitude of the sex differences in variability found in the meta-analysis is entirely in line with previous findings from both humans and non-human animals; (b) the generalized lack of statistical significance does not imply that effect sizes were too small to be considered meaningful, as the study was severely underpowered to detect effect sizes in the plausible range; (c) the results of the meta-analysis can be expected to underestimate the true magnitude of sex differences in the variability of personality, because the behavioral measures employed in most of the original studies contain large amounts of measurement error; and (d) variability effect sizes based on personality scores, latencies, and proportions suffer from lack of statistical validity, adding even more noise to the meta-analysis. In total, Harrison et al.'s study does nothing to disprove the greater male variability hypothesis in mammals, let alone in humans. To the extent that they are valid, the data remain compatible with a wide range of plausible scenarios.

Keywords: Animal personality; greater male variability; measurement error; sex differences; sexual selection.

Introduction

In their recent paper, Harrison et al. (2021) set out to test the greater male variability hypothesis with respect to personality in non-human animals (220 species of mammals, birds, reptiles/amphibians, fish, and invertebrates), along five dimensions of *activity*, *aggression*, *boldness*, *exploration*, and *sociality*. Based on the non-significant results of their meta-analysis, the authors concluded that "personality-like behavioural traits are, in general, not more variable in males than females" (p. 17) and that "there is no evidence to support the 'greater male variability' hypothesis in any taxonomic group for any of the five personality axes" (p. 18). The authors also argued that their findings have important implications for human psychology; specifically, they "raise doubts about the extent to which evolved biological differences between the sexes, arising from past sex differences in sexual and natural selection, should be used to explain why men have greater trait variation than women for certain behavioural traits" (p. 17). These conclusions were amplified in a sensationalistic press release for the study (ANU Communication & Engagement, 2021). Puzzlingly, the press release—titled *sexist "sexplanation" for men's brilliance debunked*—focused heavily on sex differences in the variability of IQ scores (which were also mentioned in the paper), despite the fact that the study was limited to animal personality and did not include any data on cognitive ability.[1]

In this commentary, we reconsider the results presented by Harrison et al. and show that their conclusions are unwarranted. Specifically, we note that (a) in mammals, birds, and reptiles/amphibians, the magnitude of the sex differences in variability found in the meta-analysis is entirely in line with previous findings from both humans and non-human animals; (b) the generalized lack of statistical significance does not imply that effect sizes were too small to be considered meaningful, as the study was severely underpowered to detect effect sizes in the plausible range; (c) the results of the meta-analysis can be expected to underestimate the true magnitude of sex differences in the variability of personality, because the behavioral measures employed in most of the original studies contain large amounts of measurement error; and (d) variability effect sizes based on personality scores, latencies, and proportions suffer from lack of statistical validity, adding even more noise to the meta-analysis.

Despite these limitations, the study by Harrison et al. makes a valuable contribution to the literature; however, it does nothing to disprove the greater male variability hypothesis in mammals, let alone in humans. To the extent that they are valid, the data are compatible with a wide range of plausible scenarios—including the possibility that, at least in some taxonomic groups, sex differences in the variability of personality may be larger than those in the variability of morphological traits. Note that, in this commentary, we focus specifically on the statistical and methodological limitations of Harrison et al.'s study, which we believe are sufficient to undermine the authors' main conclusion. We do not address other potentially relevant issues, such as the differences between personality and cognitive ability or the potential theoretical explanations of greater male variability.

[1] The authors of the study were contacted with questions, and confirmed that they had approved the press release and its focus on IQ scores (Lauren Harrison and Michael Jennions, personal communication, December 22, 2021).

Effect Sizes, Statistical Power, and the Precision of Estimates

Before researchers "embrace the null" by interpreting statistical non-significance as evidence against the existence of an effect, it is imperative that they ask if the study had sufficient power to reliably detect the effect of interest. Surprisingly, Harrison et al. did not address the issue of statistical power (or precision) at all, either in the main text or in the supporting information. The question of statistical power to detect meaningful effect sizes requires specification of a meaningful effect size. To quantify sex differences in variability, the authors used the natural logarithm of the ratio between the coefficients of variation of males and females (lnCVR); positive values indicate greater male variability, whereas negative values indicate greater female variability. They labeled the effects found in the meta-analysis as "small in magnitude" and "moderate to small" (p. 12-13), but did not provide benchmarks to support this interpretation, or compare their estimates with those reported in the rest of the literature.

Then, a crucial question remains: what is the range of plausible effect sizes for sex differences in variability? One way to answer is to consider the results of previous large-scale studies of non-human animals. In their original meta-analysis of variability in body size, Reinhold and Engqvist (2013) found significant sex differences in all the taxonomic groups, with lnCVR = 0.02 in mammals and insects, –0.02 in birds, and –0.04 in butterflies. A more sophisticated reanalysis of the same data by Nakagawa et al. (2015) yielded lnCVR = 0.04 in mammals and –0.06 in birds (all values approximated to two decimals). Wyman and Rowe (2014) found significantly greater phenotypic variability in males across species on an assortment of morphological, physiological, developmental, and behavioral traits (Table 3 in Wyman & Rowe, 2014). To quantify sex differences, these authors used the difference between the male and female coefficients of variation. The raw data of the study can be used to calculate lnCVR values, whose unweighted mean across traits and species was 0.02.[2]

Another useful source of information is the research on personality and intelligence in humans. This is especially pertinent since Harrison et al. explicitly cited the debate on human sex differences as a key motivation for their study. In these fields, the default effect size is the *variance ratio* (i.e., the ratio between the male and female variances). In the largest study of sex differences in the variability of "Big Five" personality traits in adults (Borkenau et al., 2013a), variance ratios ranged between 1.05 and 1.12 for self-reported *Extraversion*, *Openness*, *Conscientiousness*, and *Agreeableness*; the variance ratio was 0.98 for *Neuroticism*, indicating greater variability in females. Another, smaller study suggested that informant reports of personality may yield higher variance ratios than self-reports, up to about 1.20 (Borkenau et al., 2013b). Large studies of IQ scores found variance ratios of up to 1.20 in children (Arden & Plomin, 2006), and 1.13-1.19 in young adolescents (Johnson et al., 2008). The authors of the latter study also attempted to estimate the non-clinical distribution of IQ, by removing low scores that were likely due to disruptive developmental conditions. After this correction, the estimated variance ratios decreased to 1.08-1.09 (see Table 1 in Johnson et al., 2008).

Personality and intelligence are measured on arbitrary scales that lack a meaningful zero; for this reason, it does not make sense to calculate coefficients of variation or derived measures

[2] Because sample sizes were not included in the data file, we did not calculate a meta-analytic estimate or correct the effect sizes for small-sample bias (see Senior et al., 2020).

such as the lnCVR, which assume ratio-level scales. When the male and female means are equal, however, the lnCVR is identical to the lnVR or *variability ratio* (the natural logarithm of the ratio of the standard deviations; Nakagawa et al., 2015; Senior et al., 2020). This simplifying assumption allows us to make rough comparisons between the psychological data from humans and the (largely morphological) data from non-human animals. For self-reported personality traits except Neuroticism, lnVR values range from 0.02 to 0.06; the effect for Neuroticism is –0.01. Informant reports may yield values as high as 0.09. For IQ scores in adolescents (where the assumption of equal means holds to a close approximation; see Johnson et al., 2008), the lnVR values are 0.06 to 0.09 for the empirical distribution, and 0.04 for the estimated non-clinical distribution. Notably, these effect sizes are very similar to the lnCVR values found in studies of non-human animals; taken together, the available data suggest a plausible range of about 0.02 to 0.09 for traits showing greater male variability.

Crucially, it would make no sense to label these figures as "small" based on arbitrary intuitions. First, an effect size is only "large" or "small" relative to a particular research question (see Del Giudice, in press). Second, differences in variability of the magnitude described by these effect sizes can have dramatic consequences at the extremes of the distribution. Figure 1 illustrates this concept by showing male and female percentages for normally distributed traits with equal means in males and females and lnCVR (or lnVR) values of 0.02, 0.06, and 0.09. For an illustration of the same phenomenon with real-world IQ data, see Figure 2 in Johnson et al. (2008); for in-depth discussion, see Del Giudice (in press).

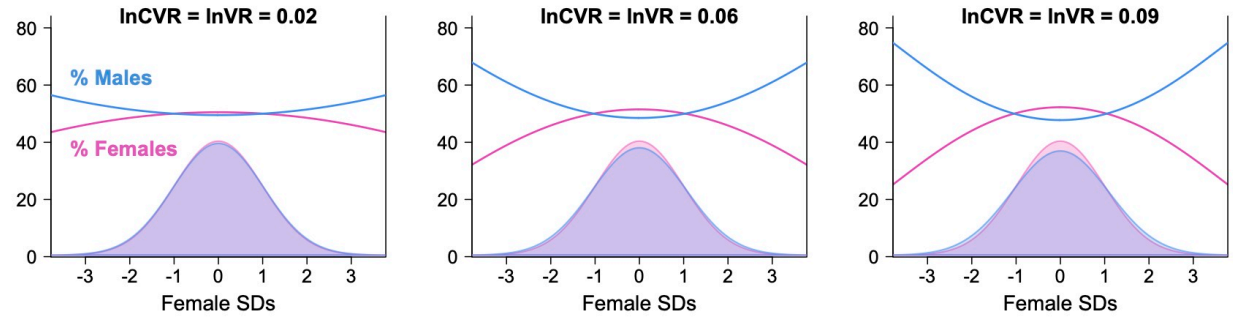

Figure 1. Percentages of males and females at different levels of a normally distributed trait with greater male variability. In these examples, the sexes have different variances but equal means, so that lnCVR = lnVR. Overlapping male and female distributions are plotted at the bottom of each plot. *Note:* the horizontal axis does not show actual trait values (which have to be positive to calculate coefficients of variation), but standard deviations from the mean in the female distribution.

Effect sizes in this range require very large samples to be reliably detected and estimated with precision. To illustrate: with $\alpha = 0.05$ (two-tailed), the sample size needed to achieve 80% power goes from about $N = 2{,}000$ for lnCVR = 0.09 to about 40,000 for lnCVR = 0.02 (see Appendix A1 for additional assumptions and details). With few exceptions, the meta-analytic estimates in the study by Harrison et al. involved dozens if not hundreds of effect sizes each. But most of those effects were based on exceedingly small samples: the supplementary data show

that the median *N* per effect size ranged from 33 in mammals to 48 in birds and invertebrates. Moreover, most studies included in the meta-analysis contributed multiple non-independent effects, reducing the effective size of the dataset. On top of this, the *p*-values reported in the main paper were adjusted upward based on the false discovery rate—a procedure that limits the proportion of "false-positive" findings, but (like all corrections for multiple testing) does so at the cost of reduced power.[3]

The validity of these concerns is fully confirmed by the meta-analytic estimates presented by Harrison et al. The overall effect sizes across personality dimensions were 0.06 in mammals, −0.14 in birds, 0.04 in reptiles and amphibians, 0.00 in fish, and −0.01 in invertebrates (Table 1 in Harrison et al.). In other words: for mammals, birds, and reptiles/amphibians, the overall magnitude of sex differences in variability *matched or exceeded* that reported in previous studies of humans and non-human animals. By definition, the fact that the corresponding tests were not statistically significant means that the study was seriously underpowered to detect effects in the plausible range. Indeed, two studies can provide very similar estimates of a certain effect, but one estimate may be "significant" and the other "non-significant", just because of differences in statistical power. Concluding that the results are inconsistent or mutually contradictory is an all-too-common fallacy in the interpretation of significance tests (see Amhrein et al., 2019); lack of statistical significance implies that the confidence interval overlaps with zero—but does not imply a zero or negligible effect size, in and of itself.

Within mammals (the taxonomic group that includes our species), the five dimensions of personality showed a consistent pattern of greater male variability, with effect sizes ranging from 0.03 to 0.11 (Table 2 in Harrison et al.). These effect size estimates were not statistically significant. But again, these results are not due to the point estimates of effect size being lower than a meaningful effect size; they were precisely in a range one could expect (indeed, somewhat larger than sex differences in the variability of morphological traits). The results are due to low power to detect effect sizes in the plausible range, because of substantial sampling variability and other likely sources of noise (see below). Table 1 in Harrison et al. reports confidence intervals around the point estimate for mammalian traits, from which one can estimate the standard errors and, thereby, power to detect a true effect of meaningful size. For true effect sizes of 0.02, 0.06, and 0.09—all within the range of plausible, meaningful effect sizes—the average power in mammals ($\alpha = 0.05$, two-tailed) was roughly 5%, 7%, and 10%, respectively (see Appendix A2).[4] In fact, the actual power was necessarily even lower, because *p*-values were adjusted for multiple testing. When power is so low, "null" effects do not allow one to be able to make any confident statement about whether true effects are meaningful, especially when point estimates are in the plausible range.

[3] In this case, most of the tests were not significant even before adjustment (see the supporting information of Harrison et al.); our point is that the costs and benefits of such adjustments should be evaluated carefully when working in conditions of low power.

[4] Note that these power estimates include *all* rejections of the null, including the probability of finding significant effects in the wrong direction (*Type S error*, with S standing for "sign"; see Gelman & Carlin, 2014). The corresponding probabilities of rejecting the null when the effect was in the right direction ranged from 3% to 9%.

The statement that power to detect effects in a plausible, meaningful range is very low is equivalent to saying that the standard errors of estimate for true effects are large relative to values in the plausible range—that is, true effect sizes are very imprecisely estimated by these data. The 95% confidence intervals presented in Table 2 obviously include values of zero; but the same intervals also include values in the 0.30-0.40 range, much greater than those typically observed for morphological traits—indeed, perhaps an order of magnitude greater than minimum values of interest. When estimates do not rule out values much larger than minimum meaningful values, it makes little sense to draw the conclusion that meaningful effects do not exist.

The large standard errors of estimate for mean lnCVR values, in turn, partly result from large standard errors for estimate of lnCVR in individual studies. Harrison et al. provided a formula for the standard error of a sample-specific estimate of lnCVR, taken from Nakagawa et al. (2015). That formula was subsequently corrected by Senior et al. (2020). We used the corrected formula to estimate that, for the 674 estimates of lnCVR in mammals, the mean standard error exceeds 0.50, which is close to the standard deviation of lnCVR estimates in mammals (see Harrison et al.'s Figure 2, which shows that > 90% of estimates fall in a vast range of −1 to 1; a lnCVR of 1 implies that the CV for males is nearly 3 times the CV for females). To be sure, some variation in lnCVR estimates across studies is due to heterogeneity in the true effect sizes; one should not expect true effect sizes to be identical across all species and across all possible measurements. But most variation, it seems, simply reflects sampling variability in individual estimates due to small sample sizes (further amplified when CVs are relatively large; see Senior et al., 2020).

The Impact of Measurement Error

By definition, personality refers to patterns of *stable individual differences* in behavior. However, actual personality measurements are also affected by other sources of variability—including within-individual differences over time and random noise—which together constitute the measurement error of personality. In the biological literature, the proportion of "true" between-individual variance over the total variance is called *repeatability*; this is conceptually equivalent to the psychometric concept of *reliability*, and is most closely approximated by *test-retest reliability* (i.e., the correlation between two administrations of the same measure at different times).

Harrison et al. noted that "[…] evidence has emerged for repeatable and heritable behavioural variation among non-human animals that is akin to human personality" (p. 3) and that "[b]y definition, personality traits are repeatable" (p. 5). But repeatability is not an all-or-none property, and can range from almost zero to almost one depending on the trait in question and the way it is assessed. When they are measured with well-validated tests, the "Big Five" of human personality show high repeatabilities, between .80 and more than .90 depending on the time interval (e.g., McCrae et al., 2011). IQ scores obtained from full-scale intelligence tests are also extremely repeatable, with test-retest reliabilities consistently over .90 (see Jensen, 1998; Roid, 2003; Wechsler, 2008). In non-human animals, aggregate ratings based on multiple behaviors can also reach high levels of repeatability (e.g., about .80 in Uher et al., 2008).

The same is *not* true of single behavioral assessments, which are strongly affected by within-individual variation and other sources of measurement error. In fact, meta-analyses of animal studies have shown that the average repeatability of single behaviors is about .30-.40 (Bell et al., 2009; Garamszegi et al., 2013); these figures mirror classical psychological findings about the consistency between single instances of behavior (see Kenrick & Funder, 1988). Note that when repeatability is less than .50, a *majority* of the variance of the behavioral measure is actually error variance, rather than variance attributable to personality. This has obvious implications when the goal is to estimate sex differences in variability: measurement error inflates the within-sex variance of both sexes, overshadowing the true differences in the variability of personality and leading to deflated estimates of the corresponding effect sizes. Figure 2 shows the relation between true and measured effect sizes at various levels of repeatability, assuming equal means in males and females (and equal amounts of measurement error in the two sexes; see Appendix A3 for details).

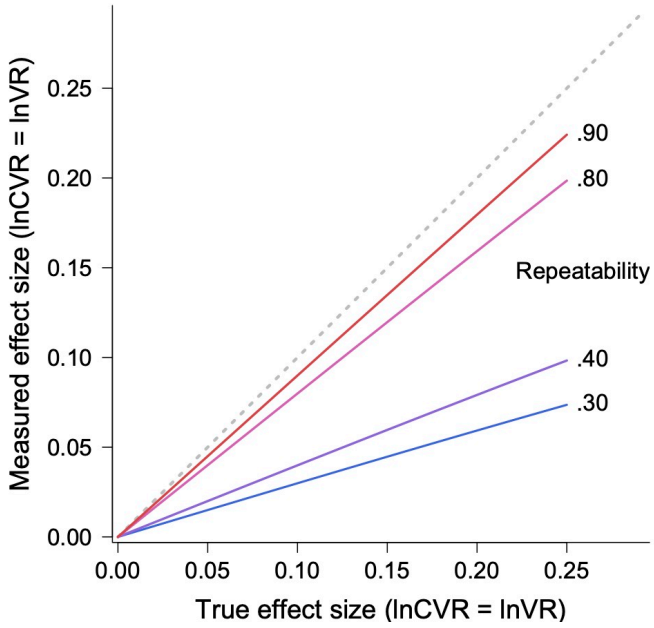

Figure 2. Expected relations between true and measured effect sizes at high vs. low levels of repeatability. The males and female distributions are assumed to have equal means, so that lnCVR = lnVR; measurement error is also assumed to have the same variance in the two sexes. Repeatabilities of .80-.90 are typical of well-validated personality and intelligence tests; values of .30-.40 are typical of single behavioral assessments.

In the meta-analysis by Harrison et al., the vast majority of the effect sizes (about 90% of the total) came from single behavioral assessments; unsurprisingly, effect sizes calculated from only behavioral assessments were virtually identical to the effect sizes calculated on the complete dataset (see Table S2 in Harrison et al.). By assuming an average repeatability of .30-.40, one can obtain a rough but realistic estimate of the true differences in variability that correspond to the measured effect sizes (see Appendix A3). For example, the observed effect size of lnCVR = 0.06 in mammals is what one would expect if the true effect size were *much*

larger, quite possibly in the 0.15-0.20 range. In contrast, morphological traits are typically measured with very high repeatability; hence, measured sex differences in morphological variability (e.g., Reinhold & Engqvist, 2013) are going to be almost identical to the corresponding true differences. These considerations underscore the fact that, at least in some taxonomic groups, Harrison et al.'s findings are compatible with sex differences in the variability of personality even larger than those in morphological variability.

Of course, the attenuating effect of measurement error on measured effect sizes contributes to reduce the power and precision of Harrison et al.'s meta-analysis. Even though in this commentary we focus specifically on sex differences in variability, measurement error has the same effect on standardized mean differences (SMD). Specifically, the measured SMD equals the true SMD multiplied by the square root of the repeatability (see Del Giudice, in press). Hence, an average repeatability of .30-.40 implies that Harrison et al.'s estimates of the SMD would correspond to true values of the SMD almost twice as large. We also note in passing that, in addition to their low repeatability, single behavioral assessments often show poor validity and consistency (e.g., Beckmann & Biro, 2013); some of the behaviors classified by the authors as indicators of personality traits such as "boldness" may not measure the relevant traits at all (see e.g., Carter et al., 2012).

To clarify, we are not claiming that Harrison et al.'s data show that true effect sizes *are* in the plausible range we identified. As noted above, the low power of their study is due to imprecision in the estimation of true effect sizes. For mammals, their estimate of lnCVR = 0.06 is a best guess for true effect size; but true effect sizes much larger and much smaller than 0.06 are also compatible with that same estimate. Our point regarding measurement error is that Harrison et al.'s estimates are not incompatible with true effects of behavioral lnCVR even larger than those observed for morphological traits—yet this conclusion is not what readers will glean from Harrison et al.'s presentation of their results.

Invalid Effect Sizes Based on Scores, Latencies, and Proportions

The final problem we want to highlight concerns the statistical validity of a subset of variability effect sizes, specifically those based on personality scores, transformed latencies, and transformed proportions. About 10% of the effect sizes included in the analysis were not based on single behavioral assessment but on personality scores. As we noted earlier, personality scores are not ratio scales, because the location of the zero of the scale is arbitrary. When the scale of a variable does not have a meaningful zero, the coefficient of variation and derived statistics such as lnCVR are meaningless and should not be calculated (Nakagawa et al., 2015; Pélabon et al., 2020). Harrison et al. ignored this caveat and calculated lnCVR values from personality scores, yielding invalid effect sizes. Fortunately, these effect sizes account for a relatively small proportion of the total; as a result, lnCVR estimates based on the complete dataset were virtually identical to those calculated excluding scores (except in invertebrates, where they were excluded from the analysis; see Table S2 in Harrison et al., 2021). Another mitigating factor is that, when the male and female means are equal, the lnCVR reduces to the lnVR, which is a meaningful statistic for personality scores if they can be treated as interval scales. In a large proportion of effect sizes, mean differences were small or close to zero (see Figures 2-6 in Harrison et al.), which probably helped reduce the impact of this error.

To normalize the distribution of right-skewed latency data (in about 9% of the effect sizes), the authors applied a log-transformation to the means and standard deviations. However, if the original variable (latency time) is measured on a ratio scale, the log-transformed variable is not, and the coefficient of variation calculated on the transformed data loses its meaning (Pélabon et al., 2020). Similarly, about 2% of the effect sizes were based on proportions (e.g., proportion of time spent hiding); in about half of the cases, the means and standard deviations were logit-transformed to normalize the distribution.[5] This transformation changes the location of the zero of the scale, so that the coefficient of variation calculated on the transformed variable becomes meaningless (see Pélabon et al., 2020). In total, about 20% of the variability effect sizes included in the meta-analysis are statistically invalid due to the improper use of the coefficient of variation. This must have added even more noise to a study that already suffered from low statistical power and high levels of measurement error.

Conclusion

Based on the non-significant findings of their study and their intuitive interpretation of effect sizes, Harrison et al. made strong claims about the supposed lack of evidence for the greater male variability hypothesis. Here we showed that the data do not support their claims. The meta-analysis was underpowered to detect effects in the plausible range; many of the estimates were entirely in line with previous findings—which is all the more interesting since those estimates had been deflated by the presence of substantial measurement error. We also pointed out that about 20% of the effect sizes included in the analysis are statistically invalid.

Taken together, these limitations cast a different light on the study by Harrison et al. Importantly, we do not mean to suggest that the study is completely uninformative. The data collected by these authors make a useful contribution to the literature, and could be reanalyzed in ways that avoid some of the problems with the original analysis. However, they must be interpreted with caution and proper consideration of their (large) margins of uncertainty. The bottom line of our commentary is that these data simply do not permit confident conclusions about the size of sex differences in variability. Not only is it a mistake to rule out meaningful effects, but the data are potentially compatible with effect sizes larger than those observed for morphological variability. The history of the greater male variability hypothesis has been incredibly long and contentious, partly due to the methodological limitations of the early research on this topic (see Feingold, 1992; Johnson et al., 2008); injecting more confusion into the debate is the last thing we need.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305-307. https://doi.org/10.1038/d41586-019-00857-9

[5] While the authors implied that the logit transformation was applied to all the proportion measures (p. 7), this was actually done in about half of the cases, as can be seen by inspecting the supplementary data of Harrison et al. (e.g., studies P115 and P204).

ANU Communication & Engagement (2021). *Sexist "sexplanation" for men's brilliance debunked*. Retrieved on January 4, 2022. https://www.anu.edu.au/news/all-news/sexist-%E2%80%9Csexplanation%E2%80%9D-for-men%E2%80%99s-brilliance-debunked

Arden, R., & Plomin, R. (2006). Sex differences in variance of intelligence across childhood. *Personality and Individual Differences, 41,* 39-48. https://doi.org/10.1016/j.paid.2005.11.027

Beckmann, C., & Biro, P. A. (2013). On the validity of a single (boldness) assay in personality research. *Ethology, 119,* 937–947. https://doi.org/10.1016/j.anbehav.2012.06.015

Bell, A. M., Hankison, S. J., & Laskowski, K. L. (2009). The repeatability of behaviour: A meta-analysis. *Animal Behaviour, 77,* 771-783. https://doi.org/10.1016/j.anbehav.2008.12.022

Borkenau, P., Hřebíčková, M., Kuppens, P., Realo, A., & Allik, J. (2013b). Sex differences in variability in personality: A study in four samples. *Journal of Personality, 81,* 49-60. https://doi.org/10.1111/j.1467-6494.2012.00784.x

Borkenau, P., McCrae, R. R., & Terracciano, A. (2013a). Do men vary more than women in personality? A study in 51 cultures. *Journal of Research in Personality, 47,* 135-144. https://doi.org/10.1016/j.jrp.2012.12.001

Carter, A. J., Marshall, H. H., Heinsohn, R., & Cowlishaw, G. (2012). How not to measure boldness: novel object and antipredator responses are not the same in wild baboons. *Animal Behaviour, 84,* 603–609. https://doi.org/10.1111/eth.12137

Del Giudice, M. (in press). Measuring sex differences and similarities. In D. P. VanderLaan & W. I. Wong (Eds.), *Gender and sexuality development: Contemporary theory and research.* Springer.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62,* 61-84. https://doi.org/10.3102/00346543062001061

Garamszegi, L. Z., Markó, G., & Herczeg, G. (2013). A meta-analysis of correlated behaviors with implications for behavioral syndromes: Relationships between particular behavioral traits. *Behavioral Ecology, 24,* 1068-1080. https://doi.org/10.1093/beheco/art033

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9,* 641-651. https://doi.org/10.1177/1745691614551642

Harrison, L. M., Noble, D. W., & Jennions, M. D. (2021). A meta-analysis of sex differences in animal personality: No evidence for the greater male variability hypothesis. *Biological Reviews.* https://doi.org/10.1111/brv.12818

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Prager.

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science, 3,* 518-531. https://doi.org/10.1111/j.1745-6924.2008.00096.x

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist, 43,* 23–34. https://doi.org/10.1037/0003-066X.43.1.23

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15,* 28-50. https://doi.org/10.1177/1088868310366253

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., & Senior, A. M. (2015). Meta-analysis of variation: Ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution, 6,* 143-152. https://doi.org/10.1111/2041-210X.12309

Pélabon, C., Hilde, C. H., Einum, S., & Gamelon, M. (2020). On the use of the coefficient of variation to quantify and compare trait variation. *Evolution Letters, 4,* 180-188. https://doi.org/10.1002/evl3.171

Reinhold, K., & Engqvist, L. (2013). The variability is in the sex chromosomes. *Evolution, 67,* 3662-3668. https://doi.org/10.1111/evo.12224

Roid, G. H. (2003). *Stanford-Binet intelligence scales, technical manual.* (5th ed.). Riverside Publishing.

Senior, A. M., Viechtbauer, W., & Nakagawa, S. (2020). Revisiting and expanding the meta-analysis of variation: The log coefficient of variation ratio. *Research Synthesis Methods, 11,* 553-567. https://doi.org/10.1002/jrsm.1423

Uher, J., & Asendorpf, J. B. (2008). Personality assessment in the Great Apes: Comparing ecologically valid behavior measures, behavior ratings, and adjective ratings. *Journal of Research in Personality, 42,* 821–838. https://doi.org/10.1016/j.jrp.2007.10.004

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale. Technical and interpretive manual* (4th ed.). Pearson.

Wyman, M. J., & Rowe, L. (2014). Male bias in distributions of additive genetic, residual, and phenotypic variances of shared traits. *The American Naturalist, 184,* 326-337. https://doi.org/10.1086/677310

Appendix

A1. Sample Size Required to Detect Sex Differences in Variability in a Single Study

Using Monte Carlo simulations, we calculated the power to detect differences in variability corresponding to lnCVR values of 0.02, 0.04, 0.06, and 0.09 in a single study, with $\alpha = 0.05$ (two-tailed). For simplicity, we assumed equal means in males and females in the population (that is, lnCVR = lnVR); the female coefficient of variation was set to 0.08, which approximates the median value for body size across species (McKellar & Hendry, 2009). The total sample size was varied from 100 to 100,000; each sample included equal numbers of males and females. Figure A1 shows the simulation results. Note that these power calculations include *all* rejections of the null, including the probability of finding significant effects in the wrong direction (*Type S error*; Gelman & Carlin, 2014). This leads to somewhat higher estimates in the low-power region of the curve, but has no effect when power is moderate to high.

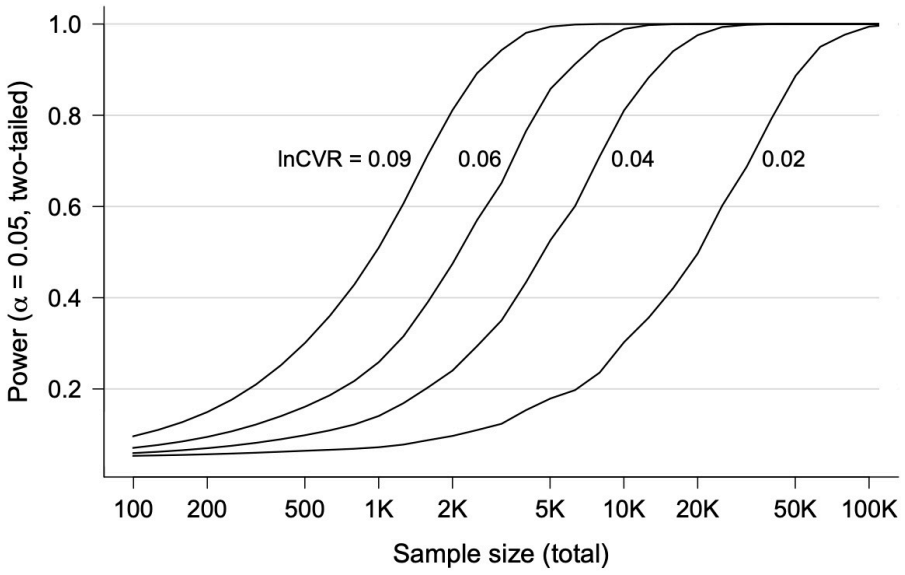

Figure A1. Power to detect differences in variability in a single study ($\alpha = 0.05$, two-tailed).

A2. Power of Harrison et al.'s Analysis in Mammals

Harrison et al.'s Table 1 reported confidence intervals for lnCVR. From the confidence interval reported for mammalian traits, we estimated the standard errors of estimate (where SE is roughly ¼ of the width of a confidence interval). The SE was estimated to be .133. Using G*power 3.1 (Faul et al., 2009) and this estimated SE, we calculated power to detect true effect sizes of .02, .06, and .09.

A3. Attenuation of Variability Effect Sizes due to Measurement Error

Repeatability is defined as the between-individual variance (V_I) divided by the total variance of the measured trait (V_T), which in turn is the sum of V_I and a residual variance term (V_R) that includes within-individual variability, random noise, etc. The repeatabilities of males and females are:

$$R_m = \frac{V_{Im}}{V_{Im} + V_{Rm}}, \; R_f = \frac{V_{If}}{V_{If} + V_{Rf}}. \quad (1)$$

If the residual variance (i.e., the measurement error) has the same magnitude in the two sexes ($V_{Rm} = V_{Rf} = V_R$), greater male variability at the between-individual level ($V_{Im} > V_{If}$) implies that males will show higher repeatability than females ($R_m > R_f$). Indeed, this is a common finding in the literature; in one meta-analysis, the effect seemed to be driven by mate choice behaviors (Bell et al., 2009), but other studies have found higher male repeatabilities in exploration, sociability, and other traits (e.g., Dingemanse et al., 2002; Strickland & Frère, 2018).

In what follows, we further assume equal means in males and females, so that lnCVR = lnVR. With an even sex ratio, the repeatability R calculated on the population as a whole is:

$$R = \frac{\frac{V_{Im} + V_{If}}{2}}{\frac{V_{Im} + V_{If}}{2} + V_R}. \quad (2)$$

Rearranging Eq. 2 yields the residual variance V_R:

$$V_R = \frac{V_{Im} + V_{If}}{2} \cdot \frac{1-R}{R}. \quad (3)$$

By definition, the true lnVR (in the population) is:

$$\text{lnVR} = \ln\sqrt{\frac{V_{Im}}{V_{If}}}, \quad (4)$$

while the measured lnVR* is:

$$\text{lnVR}^* = \ln\sqrt{\frac{V_{Tm}}{V_{Tf}}} = \ln\sqrt{\frac{V_{Im} + V_R}{V_{If} + V_R}}. \quad (5)$$

Define λ as the true variance ratio V_{Im}/V_{If}, so that $V_{Im} = \lambda V_{If}$. From Eq. 4,

$$\lambda = e^{2 \cdot \text{lnVR}}. \quad (6)$$

Substituting Eq. 3 into Eq. 5 yields:

$$\ln VR^{*} = \ln \sqrt{\frac{V_{Im} + \frac{V_{Im}+V_{If}}{2} \cdot \frac{1-R}{R}}{V_{If} + \frac{V_{Im}+V_{If}}{2} \cdot \frac{1-R}{R}}} = \ln \sqrt{\frac{\lambda + \frac{\lambda+1}{2} \cdot \frac{1-R}{R}}{1 + \frac{\lambda+1}{2} \cdot \frac{1-R}{R}}} , \quad (7)$$

which, together with Eq. 6, can be used to calculate the measured $\ln VR^{*}$ from the true lnVR and the repeatability *R*.

A4. References

Bell, A. M., Hankison, S. J., & Laskowski, K. L. (2009). The repeatability of behaviour: A meta-analysis. *Animal Behaviour, 77,* 771-783. https://doi.org/10.1016/j.anbehav.2008.12.022

Dingemanse, N. J., Both, C., Drent, P. J., Van Oers, K., & Van Noordwijk, A. J. (2002). Repeatability and heritability of exploratory behaviour in great tits from the wild. *Animal Behaviour, 64*, 929-938. https://doi.org/10.1006/anbe.2002.2006

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9,* 641-651. https://doi.org/10.1177/1745691614551642

McKellar, A. E., & Hendry, A. P. (2009). How humans differ from other animals in their levels of morphological variation. *PLoS ONE, 4,* e6876. https://doi.org/10.1371/journal.pone.0006876

Strickland, K., & Frère, C. H. (2018). Predictable males and unpredictable females: repeatability of sociability in eastern water dragons. *Behavioral Ecology, 29,* 236-243. https://doi.org/10.1093/beheco/arx148