The S-Index:
Summarizing Patterns of Sex Differences at the Distribution Extremes

Marco Del Giudice

University of New Mexico

In press (2023): ***Personality and Individual Differences***

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

Sex differences researchers are becoming increasingly interested in how differences in averages and variability jointly influence male and female representation at the tails of the distribution. This paper introduces the *S-index*, a novel index that provides a simple and robust summary of the shape of sex differences at the distribution extremes. The use of *S* is illustrated with a selection of real-world datasets of personality and cognitive ability, and a R function is provided to calculate *S* and draw intuitive proportion plots of sex differences across the distribution. The S-index is not limited to the study of sex differences; it can be applied to other domains as long as the groups to be compared are about equally represented in the population and the variables of interest are approximately bell-shaped.

Keywords: gender differences; group differences; methodology; proportions; sex differences; tails.

In the psychological literature, sex differences are typically measured by comparing male and female averages with univariate indices, such as Cohen's *d* (e.g., Hyde, 2014).[1] In recent years, researchers have started exploring alternative methods to reveal the full scope and complexity of sex-related patterns of cognition and behavior. These methods range from straightforward extensions of the standard approach (e.g., multivariate distances between means; see Del Giudice, 2009, 2022; Eagly & Revelle, 2022) to complex machine learning models for classification and prediction (e.g., Loesche, 2019). Meanwhile, the long-standing focus on averages has been broadened by a renewed interest in sex differences in variability, typically quantified as *variance ratios* (e.g., Borkenau et al., 2013; Gray et al., 2019; Johnson et al., 2008; see Del Giudice, 2022). Patterns of dispersion around the mean are especially relevant to the *greater male variability hypothesis* (GMVH), which was proposed by Darwin (1871) and went on to became the object of more than a century of heated debate (see Del Giudice, 2023; Feingold, 1992). More recently, the GMVH has attracted the attention of evolutionary psychologists (e.g., Archer & Mehdikhani, 2003; Del Giudice et al., 2018; Stewart-Williams & Halsey, 2021), and shown an undiminished power to excite controversy (for a recent example, see Harrison et al., 2022 and Del Giudice & Gangestad, 2022).

Importantly, sex differences often become magnified as one moves toward more extreme values of a trait. The same difference may have a small or negligible impact near the center of the distribution, but yield substantial imbalances in the proportions of males and females at the distribution's tails. Moreover, the effects of differences in trait means and variability—as well as other features, e.g., subtle differences in skewness and kurtosis—combine with one another in ways that are not always intuitive. Patterns of sex differences at the extremes are an important topic of investigation in their own respect; for two notable empirical examples, see Paessler's (2015) study of vocational interests and Thöni and Volk's (2021) analysis of time, risk, and social preferences.

At present, researchers have two main ways to quantify sex differences at the distribution extremes. *Tail ratios* measure the ratio of the two sexes in the region above or below a cutoff (see Del Giudice, 2022; Voracek et al., 2013). *Relative distribution methods* can be used to plot the relative density of males and females across the entire distribution, and separate the effect of mean differences from that of differences in dispersion and other aspects of distribution shape (Handcock & Morris, 1998, 1999; see Del Giudice, 2022). Because ratios tend to amplify the impact of fluctuations in the data, tail ratios are quite sensitive to sampling error and require very large *N*s to yield accurate estimates. While relative density plots can be highly informative, they are not immediately intuitive and require some technical background to interpret correctly. Methods based on differences between distribution quantiles (Rousselet et al., 2017) yield results that are potentially more robust, but even harder to interpret for casual readers.

In this paper, I introduce the *S-index*, a novel statistic that provides a simple, robust summary of the shape of group differences at the distribution extremes. The S-index is designed

[1] Here I use "sex differences" as a descriptive label for differences between males and females, with no particular assumptions about their biological and/or cultural origins. For more discussion of this terminological issue see Del Giudice (2022, 2023).

to compare groups that are about equally represented in the population, which makes it ideally suited to the analysis of on sex differences. In what follows, I describe the logic and meaning of *S* and demonstrate its use in a selection of real-world datasets. I also provide an R function to calculate the index and supplement it with intuitive, informative plots.

The S-Index

The S-index (for "shape") is based on the proportions of two groups at three points of their overall distribution: the unweighted group mean and the upper and lower extremes of the range under consideration. Proportions are considerably more robust than ratios, as they compress the upper and lower ends of the scale into a small interval (for example, increasing the M:F ratio from 20 to 200 to 2,000 only changes the percentage of males from 95% to 99.5% to 99.9%). For this reason, proportions are not just intuitive but also comparatively insensitive to sampling error. Male and female proportions across the distribution can be displayed in *proportion plots* like the ones shown in Figure 1 (which are based on idealized normal distributions with different means and SDs).

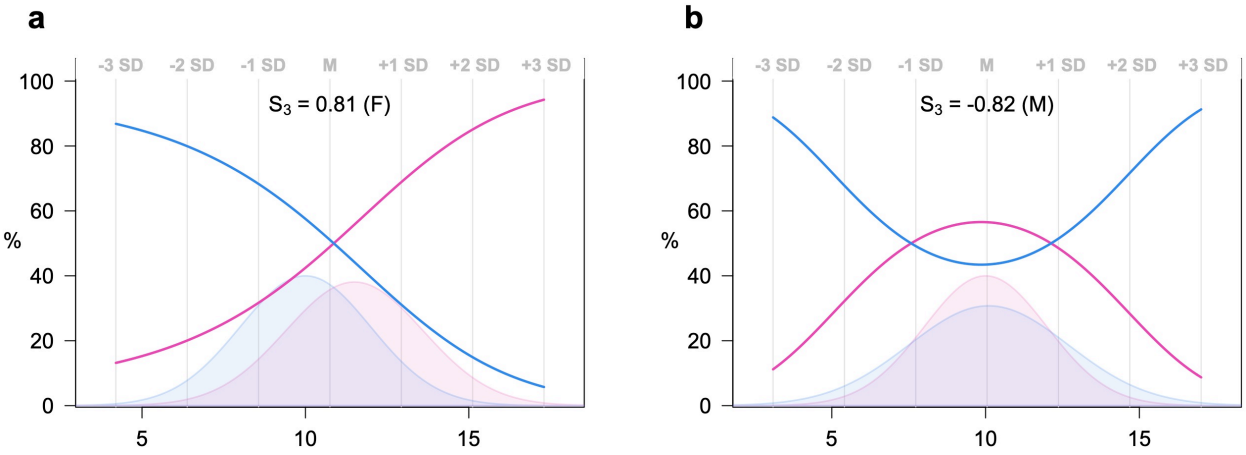

Figure 1. Illustrative proportion plots showing (a) an X-shaped pattern and (b) a U-shaped pattern of sex differences, based on idealized normal distributions. Blue = males; pink = females.

When two groups are about equally represented in the population (as is the case for males and females) and their distributions are approximately bell-shaped (as with many morphological and psychological traits), the resulting patterns of proportions fall between two prototypical configurations. The first prototype corresponds to X-shaped patterns like the one in Figure 1a: the proportion of members of one group increases moving from the lower to the upper extreme of the distribution, while the proportion of members of the other group declines accordingly. The second prototype corresponds to U-shaped patterns line that in Figure 1b: the proportion of members of one group increases at both extremes and decreases around the mean (U), while the proportion of members of the other group declines at both extremes and increases toward the mean (inverted U). In between one finds weaker versions of X-shaped and U-shaped patterns, as well as "mixed" patterns that combine aspects of the two. The online supplement

(section S1) shows the range of patterns that may arise from combinations of mean and variance differences typically observed for psychological traits.

The S-index exploits the regularities of proportion patterns to summarize their shape with a single numerical value ranging from –1 to +1. Positive values of *S* correspond to X-shaped patterns (e.g., Fig. 1a), whereas negative values correspond to U-shaped patterns (e.g., Fig. 1b). Large values of *S* indicate strong patterns, such as those shown in Fig. 1; smaller values indicate weaker and/or more mixed patterns. Finally, values around zero mean that the pattern lacks a consistent trend on at least one side of the distribution, possibly because there are no meaningful differences in the trait. The absolute magnitude of *S* can be used to rank different effects as weaker or stronger than one another, but should not be over-interpreted as implying a true ratio scaling. For example, an effect of $S = -.60$ indicates a more pronounced U-shaped pattern than one of $S = -.30$; however, that pattern may not be exactly "twice as large" as the other on any straightforward metric.

Formula and Notation

The formula to compute *S* is:

$$S = \operatorname{sgn}[(p_{\mathrm{L}} - p_{\mathrm{M}})(p_{\mathrm{M}} - p_{\mathrm{U}})] \cdot \sqrt{\frac{|p_{\mathrm{L}} - p_{\mathrm{M}}|}{\frac{1}{2} + \left(\frac{1}{2} - p_{\mathrm{M}}\right)\operatorname{sgn}(p_{\mathrm{L}} - p_{\mathrm{M}})} \cdot \frac{|p_{\mathrm{U}} - p_{\mathrm{M}}|}{\frac{1}{2} + \left(\frac{1}{2} - p_{\mathrm{M}}\right)\operatorname{sgn}(p_{\mathrm{U}} - p_{\mathrm{M}})}}$$

where $\operatorname{sgn}(\cdot)$ is the sign function and p_{M}, p_{U}, and p_{L} are the proportions (range: 0-1) of one of the groups (which one does not matter) at the unweighted mean of the two groups and at the upper and lower extremes of the range under consideration, respectively (see Fig. 2). The square-root term in the formula is the geometric mean of two normalized differences: (1) the absolute difference between p_{M} and p_{L}, normalized by either p_{M} or $(1 - p_{\mathrm{M}})$, depending on the side of p_{M} on which p_{L} lies; and (2) the absolute difference between p_{M} and p_{U}, also normalized by p_{M} or $(1 - p_{\mathrm{M}})$ depending on the value of p_{U}. In the example of Fig. 2, $|p_{\mathrm{L}} - p_{\mathrm{M}}|$ would be normalized by p_{M}, whereas $|p_{\mathrm{U}} - p_{\mathrm{M}}|$ would be normalized by $(1 - p_{\mathrm{M}})$. The term outside the square root determines the sign of *S*, based on the concordant vs. discordant directions of the two differences. The formula yields $S = 1$ when $p_{\mathrm{L}} = 0$ and $p_{\mathrm{U}} = 1$ or vice versa (extreme X-shaped pattern); $S = -1$ when $p_{\mathrm{L}} = p_{\mathrm{U}} = 0$ or $p_{\mathrm{L}} = p_{\mathrm{U}} = 1$ (extreme U-shaped pattern); and $S = 0$ when $p_{\mathrm{L}} = p_{\mathrm{M}}$, $p_{\mathrm{U}} = p_{\mathrm{M}}$, or both (inconsistent trends or no group differences).

The notation of *S* includes a subscript to indicate the distance of the extremes from the mean, expressed in terms of the overall standard deviation of the two groups combined (assuming equal group size). For example, S_2 would indicate the S-index calculated ±2 SDs from the mean. This raises the issue of what is a suitable operational definition of "extremes". I propose that S_3 should serve as a sensible default for most applications, keeping in mind that any such choice is going to be somewhat arbitrary. Typical rules of thumb to identify outliers use thresholds of ±2.5 or 3.5 SDs from the mean, suggesting that deviations of ±3 SDs match most researchers' intuitive conception of "extreme scores" reasonably well. In a normal distribution, ±3 SDs leave out the top and bottom 0.13% of the data, consistent with a notion of rarity at the

extremes (only about 2-3 participants out of 1,000 are expected to score outside these bounds). That said, different applications may suggest different definitions of the distribution extremes. For example, S_2 may be more informative or relevant than S_3 for certain specific purposes; in general, however, ±2 SDs from the mean identify scores that are far from average but not truly extreme (e.g., almost 5% of the data in a normal distribution are expected to lie outside these bounds).

Because S has the same value regardless of which group is used to compute p_M, p_L, and p_U, it can be useful extend the notation to specify the directionality of the pattern. A simple way to do so is to add the letter M or F to indicate whether males or females are over-represented at the upper extreme of the distribution. To illustrate, the X-shaped pattern in Fig. 1a (with more females than males at the upper extreme) can be summarized as S_3 = .81 (F), whereas the U-shaped pattern of Fig. 1b (with more males than females at both extremes) corresponds to S_3 = –.82 (M).

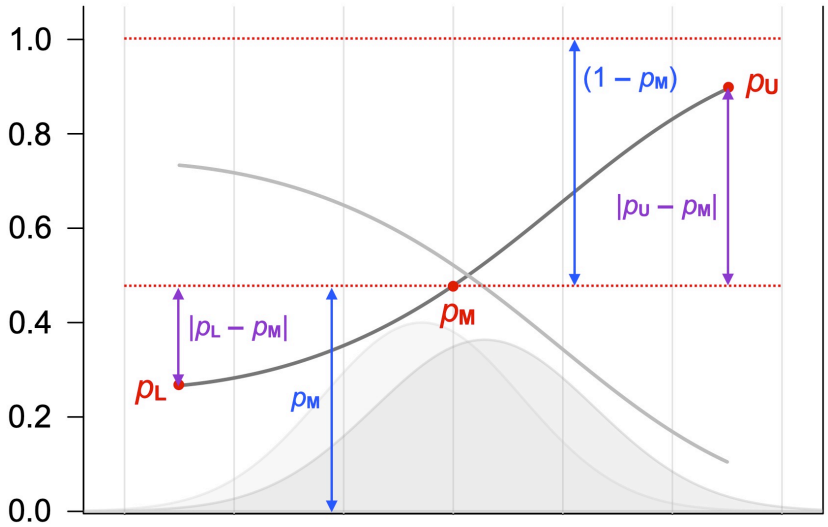

Figure 2. Visual explanation of the quantities used to calculate the S-index. The labels p_M, p_U, and p_L indicate the proportions of one of the groups (darker line) at the population mean and at the upper and lower extreme of the range under consideration.

R Function

An R function (*s.index*) to calculate the S-index and draw proportions plots is available at https://doi.org/10.6084/m9.figshare.21437727. This function can calculate the S-index from the means and SDs of two normal distributions, from the empirical distributions of two groups, or from idealized normal distributions with the same means and SDs as the empirical distributions. In the case of empirical data, density functions and proportions are estimated with a Gaussian kernel, using the average of the bandwidths suggested for the two groups by Silverman's rule of thumb (Silverman, 1986). The bandwidth used to estimate proportions is increased by an adjustment factor (1.5 by default) to yield smoother, more robust curves. The function also provides bootstrapped confidence intervals on *S*.

Statistical Power, Estimation Bias, and Standard Error

The online supplement reports initial simulation results on the inferential properties of the estimation/bootstrap procedure employed by the *s.index* function in relation to effect and sample size. These simulations explore index S_3 in the basic scenario of normally distributed populations. Section S2 displays the results for statistical power, Section S3 for estimation bias, and Section S4 for the standard error; here I present some illustrative highlights.

As can be expected, the S-index requires fairly large samples in order to reliably detect patterns of group differences at the extremes. To detect clear-cut effects of S_3 = +/–.50 with sufficient power (> 80%), one needs more than ~3,000 participants per group; achieving the same power for subtler effects around S_3 = +/–.25 requires more than ~20,000-25,000 participants per group. With ~100,000 participants per group, one obtains excellent power for all but the most subtle effects (down to about S_3 = +/–.15). Conversely, ~1,000 participants per group are only adequate to detect large effects in the order of S_3 = +/–.70.

As shown in section S3, sample estimates of S are biased toward zero, yielding somewhat deflated values of the statistic. This reflects the increasing asymmetry of the sampling distribution as sample size decreases and/or the population value of S approaches +/–1 (see sections S3 and S4). The effect becomes especially pronounced when there are fewer than ~1000 participants per group. Having more than ~10,000 participants per group reduces bias to uniformly low amounts (less than about +/–0.05 in absolute magnitude). In general, one should be very cautious about estimating S_3 with fewer than 1,000 participants per group, unless the expected effect is extremely strong (in excess of +/–.80) and significant amounts of deflation can be tolerated.

Empirical Examples

Figure 3 illustrates the S-index with a selection of real-world datasets (the data and R code used to make the plots are available at https://doi.org/10.6084/m9.figshare.21437949). Panel 3a displays the distribution of height for 7,424 adults (50.7% females) in the National Health and Nutrition Examination Survey (NHANES) database. Height is a useful reference trait because of its familiarity; sex differences show a very strong X-shaped pattern, with S_3 = 1.00 (M, 95% CI [1.00, 1.00]). When the S-index is computed from empirical datasets, the default option is to avoid extrapolating beyond the actual data: if the minimum ad/or maximum observed values of the variable fall *inside* the specified bounds (e.g., ± 3 SDs from the mean), the empirical minimum and/or maximum are used instead. Proportion plots show the empirical minimum and maximum of the variable as dotted vertical lines.

Panels 3b-3f are based on Big Five personality traits for a sample of 100,000 participants from the United States (50% females), randomly selected from a larger dataset described in Kaiser (2019). Both Agreeableness and Neuroticism yield clear X-shaped patterns, with S_3 = .56 (F, 95% CI [.47, .64]) for Agreeableness and S_3 = .47 (F, 95% CI [.37, .56]) for Neuroticism. Openness and Extraversion show much weaker (and less symmetric) configurations that do not reach statistical significance, with S_3 = .10 (F), 95% CI [–.28, .32] and S_3 = .15 (F), 95% CI [–

.19, .29], respectively. Note that these small values hide a higher proportion of males at the low end of the scale—a reminder that *S* provides a broad-brush statistical summary that cannot replace a detailed examination of the data. Finally, the S-index indicates no consistent patterns for Conscientiousness (S_3 = –.06, M, 95% CI [–.11, .05]).

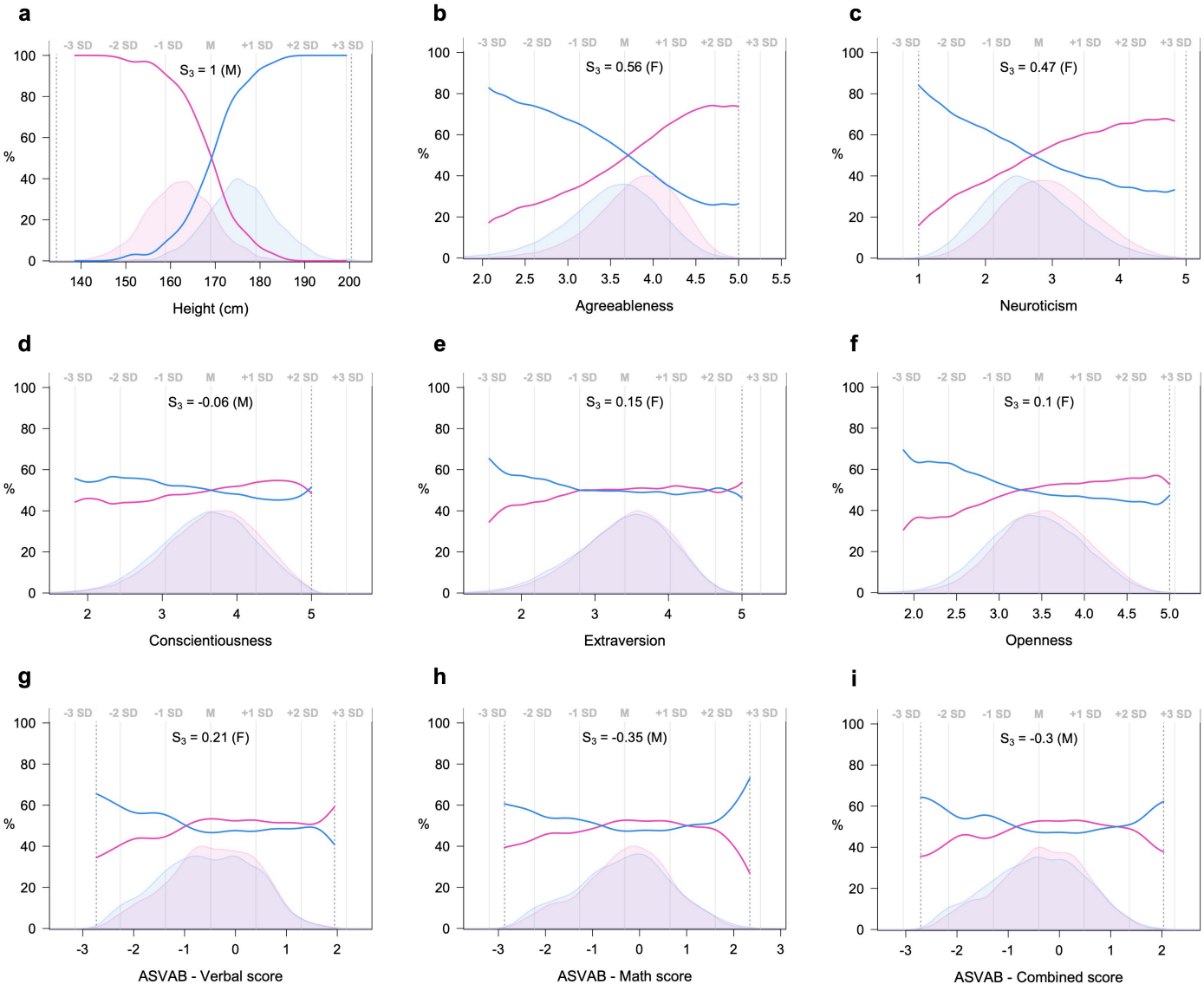

Figure 3. Examples of proportion plots and S_3 values from three empirical datasets (see main text for details). (a) Sex differences in adult height. (b-f) Sex differences in Big Five personality traits. (g-i) Sex differences in cognitive performance (ASVAB battery). Dotted vertical lines show the minimum/maximum observed values of the variable. Blue = males; pink = females.

Panels 3g-3i show the distribution of cognitive performance on the Armed Services Vocational Aptitude Battery (ASVAB). This sample from the 1997 National Longitudinal Survey of Youth (NLSY97) includes 7,076 participants with complete data (49.4% females).

Verbal scores yield a mild X-shaped pattern of sex differences, with S_3 = .21 (F, 95% CI [–.22, .40]). In contrast, math scores reveal a reliable U-shaped pattern, with a preponderance of males at the higher *and* lower ability levels: S_3 = –.35 (M, 95% CI [–.49, –.22]). The combined scores (average of math and verbal) show the same configuration (S_3 = –.30, M; 95% CI [–.44, –.08]). Given the comparatively small size of the ASVAB sample, one should keep in mind that these estimates are likely to be somewhat deflated (see supplementary section S3).

Conclusion

The S-index extends the statistical toolkit for the study of sex differences, by providing a simple and robust summary of patterns of male and female proportions at the distribution extremes. This novel index can be used to screen variables for potentially interesting patterns, and permits quantitative evaluations and comparisons in addition to qualitative ones. I have introduced the S-index in the context of psychological sex differences because its assumptions (approximately equal-sized groups, bell-shaped distributions) are often satisfied in this area of research. However, it is worth stressing that this index is by no means specific to sex differences, and can be applied equally well to other domains as long as the relevant assumptions are met.

References

Archer, J. & Mehdikhani, M. (2003). Variability among males in sexually-selected attributes. *Review of General Psychology, 7,* 219–236. https://doi.org/10.1037/1089-2680.7.3.219

Borkenau, P., McCrae, R. R., & Terracciano, A. (2013). 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

Darwin, C. (1871). *The descent of man, and selection in relation to sex*. John Murray.

Del Giudice, M. (2009). On the real magnitude of psychological sex differences. *Evolutionary Psychology, 7,* 264-279. https://doi.org/10.1177/147470490900700209

Del Giudice, M. (2022). Measuring sex differences and similarities. In D. P. VanderLaan & W. I. Wong (Eds.), *Gender and sexuality development: Contemporary theory and research*. Springer. https://doi.org/10.1007/978-3-030-84273-4_1

Del Giudice, M. (2023). Ideological bias in the psychology of sex and gender. In C. L. Frisby, W. T. O'Donohue, R. E. Redding, & S. O. Lilienfeld (Eds)., *Political bias in psychology: Nature, scope, and solutions.* Springer.

Del Giudice, M., Barrett, E. S., Belsky, J., Hartman, S., Martel, M. M., Sangenstedt, S., & Kuzawa, C. W. (2018). Individual differences in developmental plasticity: A role for early androgens? *Psychoneuroendocrinology, 90,* 165-173. https://doi.org/10.1016/j.psyneuen.2018.02.025

Del Giudice, M., & Gangestad, S. W. (2022). No evidence against the greater male variability hypothesis: A commentary on Harrison et al.'s meta-analysis of animal personality. *PsyArXiv*, https://doi.org/10.31234/osf.io/6ua8r

Eagly, A. H., & Revelle, W. (2022). Understanding the magnitude of psychological differences between women and men requires seeing the forest and the trees. *Perspectives on Psychological Science, 17,* 1339–1358. https://doi.org/10.1177/17456916211046006

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

Gray, H., Lyth, A., McKenna, C., Stothard, S., Tymms, P., & Copping, L. (2019). Sex differences in variability across nations in reading, mathematics and science: A meta-analytic extension of Baye and Monseur (2016). Large-*scale Assessments in Education, 7,* 1-29. https://doi.org/10.1186/s40536-019-0070-9

Handcock, M. S., & Morris, M. (1998). Relative distribution methods. *Sociological Methodology, 28,* 53–97. https://doi.org/10.1111/0081-1750.00042

Handcock, M. S., & Morris, M. (1999). *Relative distribution methods in the social sciences.* Springer.

Harrison, L. M., Noble, D. W., & Jennions, M. D. (2022). A meta-analysis of sex differences in animal personality: No evidence for the greater male variability hypothesis. *Biological Reviews, 97,* 679-707. https://doi.org/10.1111/brv.12818

Hyde, J. S. (2014). Gender similarities and differences. Annual Review of Psychology, 65, 373-398. https://doi.org/10.1146/annurev-psych-010213-115057

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

Kaiser, T. (2019). Nature and evoked culture: Sex differences in personality are uniquely correlated with ecological stress. *Personality and Individual Differences, 148,* 67–72. https://doi.org/10.1016/j.paid.2019.05.011

Loesche, P. M. (2019). Estimating the true extent of gender differences in scholastic achievement: A neural network approach. *Intelligence, 77,* 101398. https://doi.org/10.1016/j.intell.2019.101398

Paessler, K. (2015). Sex differences in variability in vocational interests: Evidence from two large samples. *European Journal of Personality, 29,* 568-578. https://doi.org/10.1002/per.2010

Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2017). Beyond differences in means: Robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience, 46,* 1738-1748. https://doi.org/10.1111/ejn.13610

Silverman, B. W. (1986). *Density Estimation*. Chapman & Hall.

Stewart-Williams, S., & Halsey, L. G. (2021). Men, women and STEM: Why the differences and what should be done? *European Journal of Personality, 35,* 3-39. https://doi.org/10.1177/0890207020962326

Thöni, C., & Volk, S. (2021). Converging evidence for greater male variability in time, risk, and social preferences. *Proceedings of the National Academy of Sciences, 118,* e2026112118. https://doi.org/10.1073/pnas.2026112118

Voracek, M., Mohr, E., & Hagmann, M. (2013). On the importance of tail ratios for psychological science. *Psychological Reports, 112,* 872-886. https://doi.org/10.2466/03.PR0.112.3.872-886

Supplementary material for

The S-Index: Summarizing Patterns of Sex Differences at the Distribution Extremes

S1. Patterns of proportions

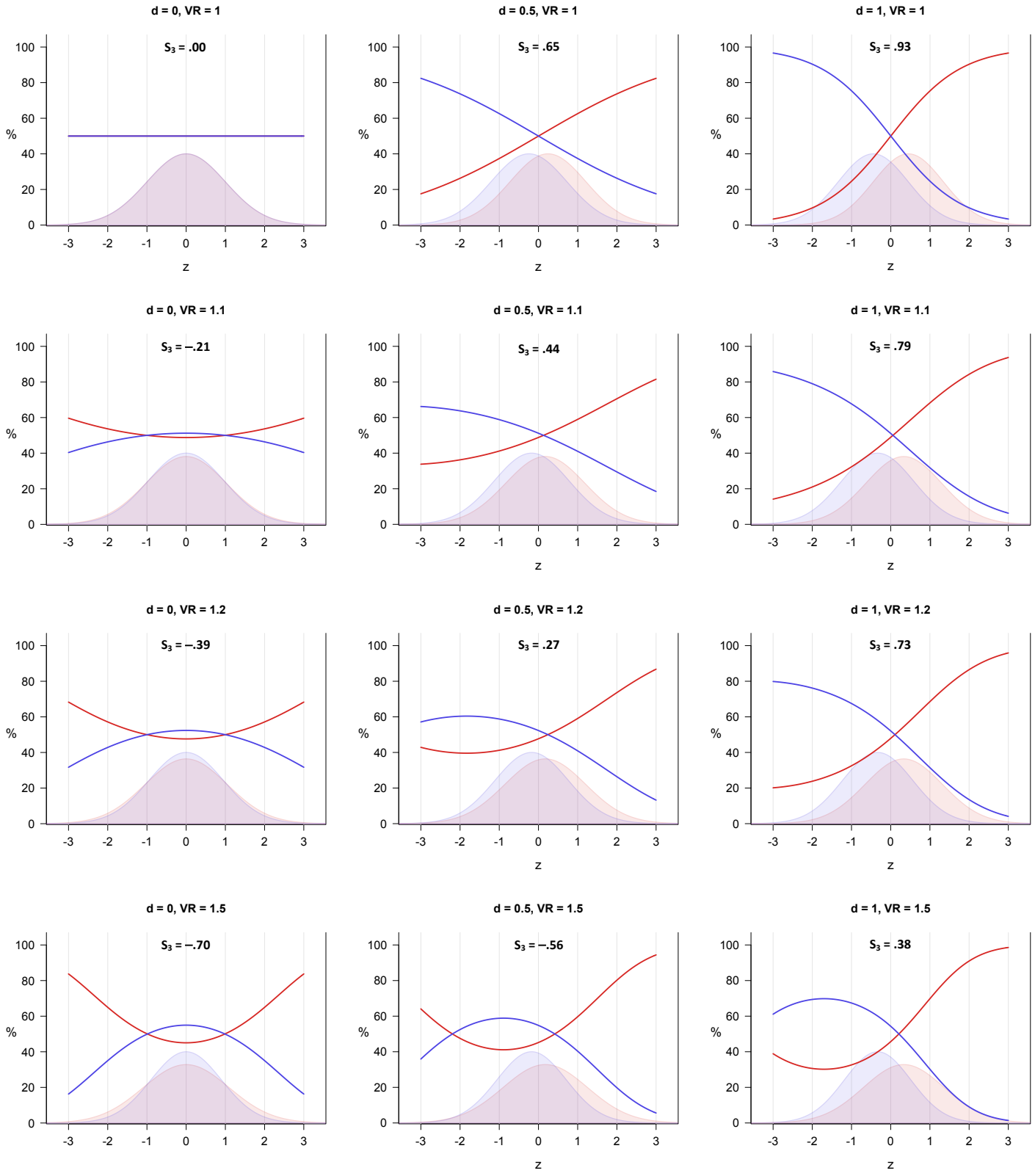

Fig. S1. Patterns of proportions (between ± 3 SDs from the mean) for two normally distributed groups of equal size. Columns correspond to increasing values of Cohen's *d* (0, 0.5, 1); rows correspond to increasing values of the variance ratio (*VR*; 1, 1.1., 1.2, 1.5).

S2. Statistical Power

The power curves displayed in Fig. S2 and S3 were obtained by (a) simulating samples from normal distributions and (b) estimating the S-index (specifically S_3) and its 95% CI from empirical data with the default settings of function *s.index* (to speed up computations, bootstrapped CIS were based on 500 samples without jackknife acceleration). Positive S_3 values were obtained from two populations with equal variances but different means, as in the top row of Fig. S1. Negative S_3 values were obtained from two populations with equal means but different variances, as in the left column of Fig. S1. In both cases, power was evaluated at 9 effect sizes and 13 sample sizes, from N = 100 to 100K per group (total sample size: 200 to 200K). The curves were based on 200 samples for each combination of effect/sample size and smoothed for error correction.

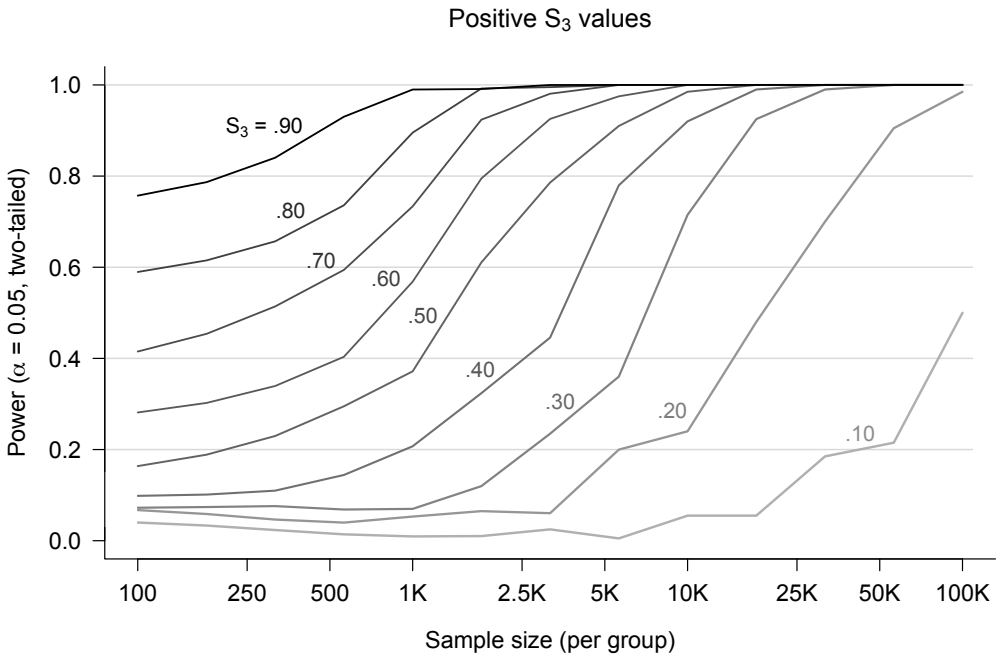

Fig. S2. Statistical power (α = 0.05, two-tailed) for positive values of the population S_3, based on equal-sized samples from two normally distributed populations.

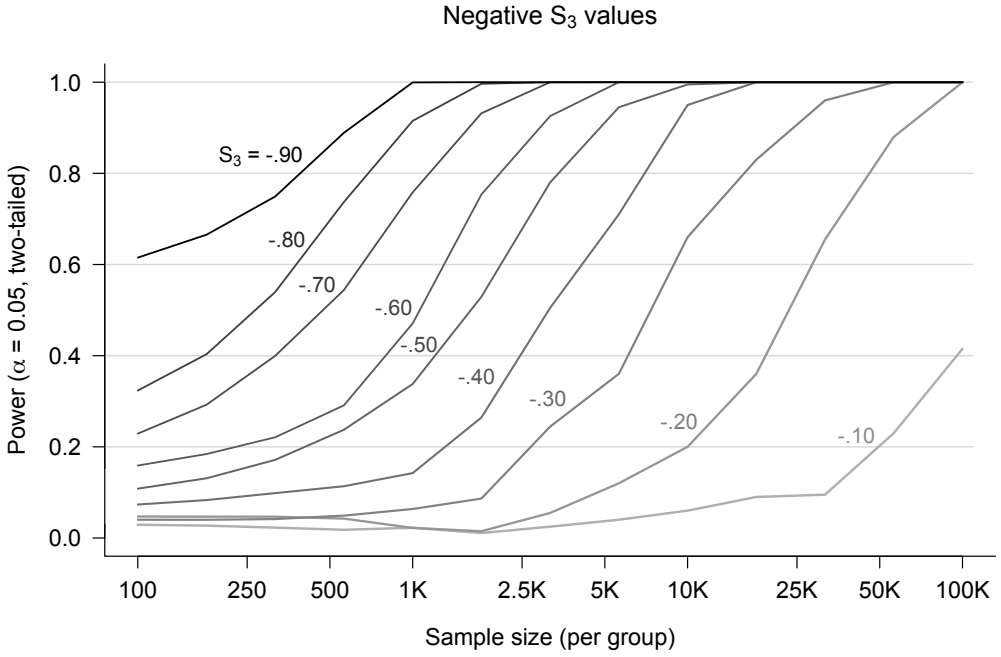

Fig. S3. Statistical power (α = 0.05, two-tailed) for negative values of the population S_3, based on equal-sized samples from two normally distributed populations.

S3. Estimation Bias

The bias curves displayed in Fig. S4 and S5 were obtained from the same simulated data presented in section S2. The curves show the average S_3 estimated with function *s.index* (smoothed for error correction), as compared with the true value of S_3 in the population.

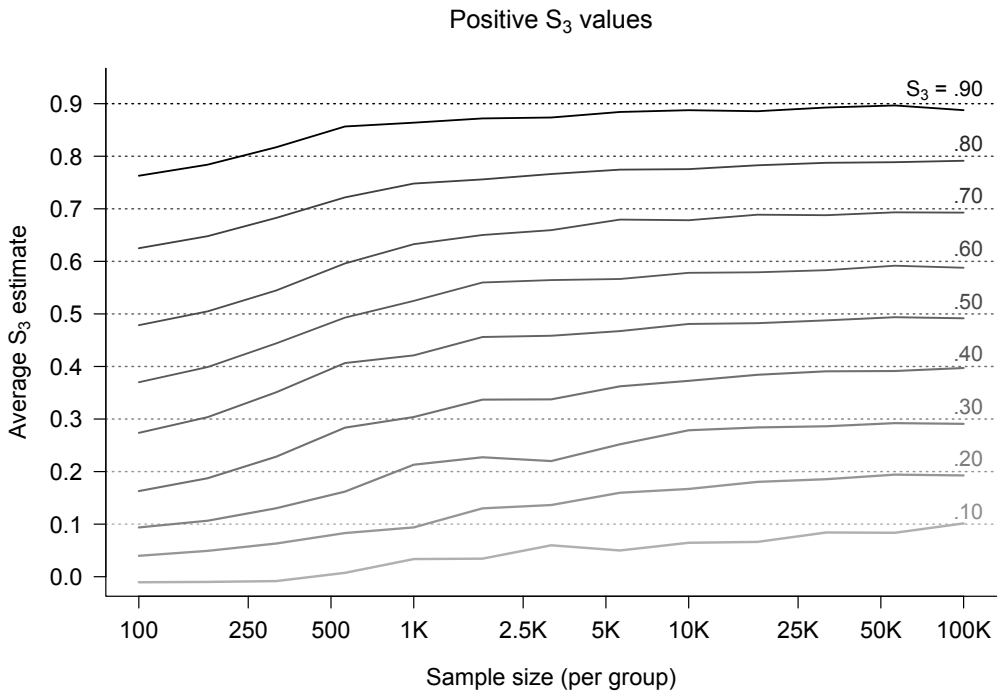

Fig. S4. Average sample estimates for positive values of S_3 at different sample sizes. Dotted lines show population values of S_3. Simulations were based on equal-sized samples from two normally distributed populations.

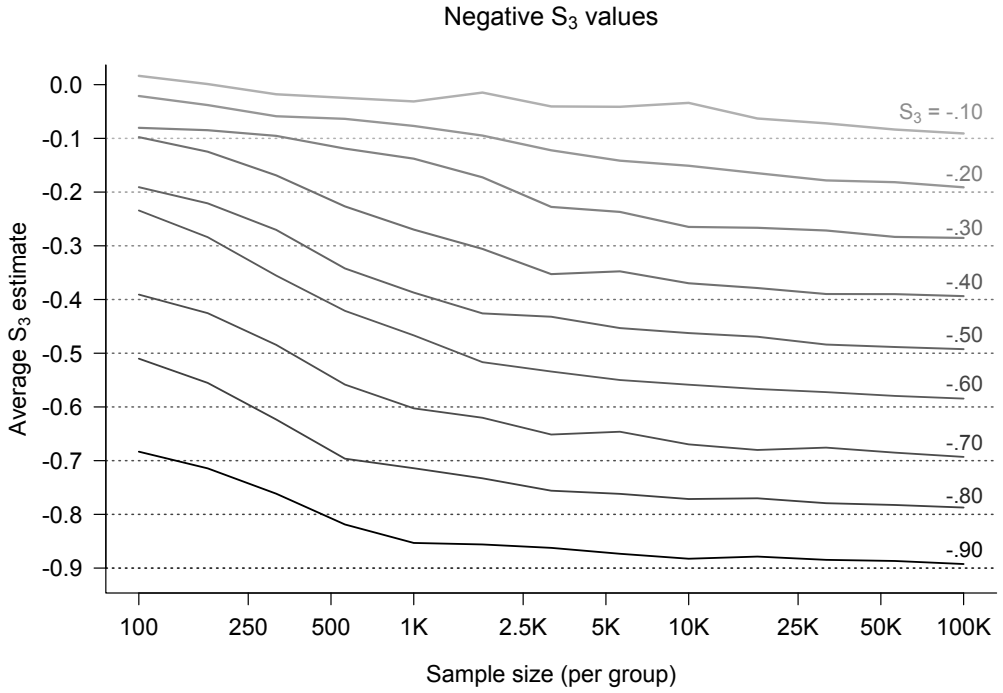

Fig. S5. Average sample estimates for negative values of S_3 at different sample sizes. Dotted lines show population values of S_3. Simulations were based on equal-sized samples from two normally distributed populations.

S4. Standard Error

The curves displayed in Fig. S6 and S7 were obtained with the same procedure as in sections S2 and S3, but with 1,000 samples for each combination of effect/sample size. The standard error was estimated as the SD of the distribution of sample values of S_3, calculated with the default settings of function *s.index*.

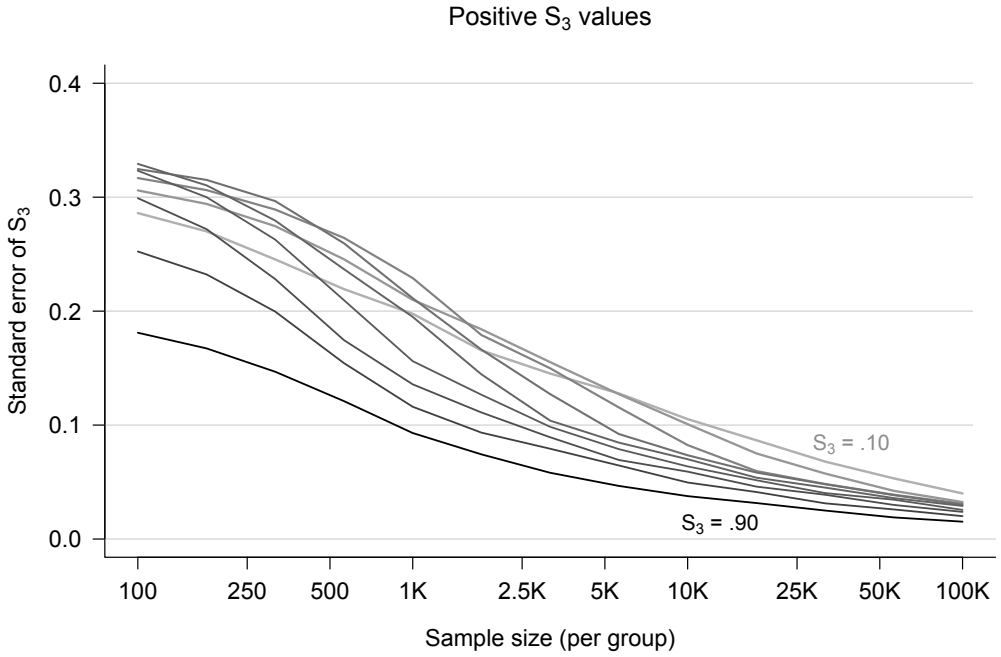

Fig. S6. Standard error estimates for positive values of the population S_3, based on equal-sized samples from two normally distributed populations.

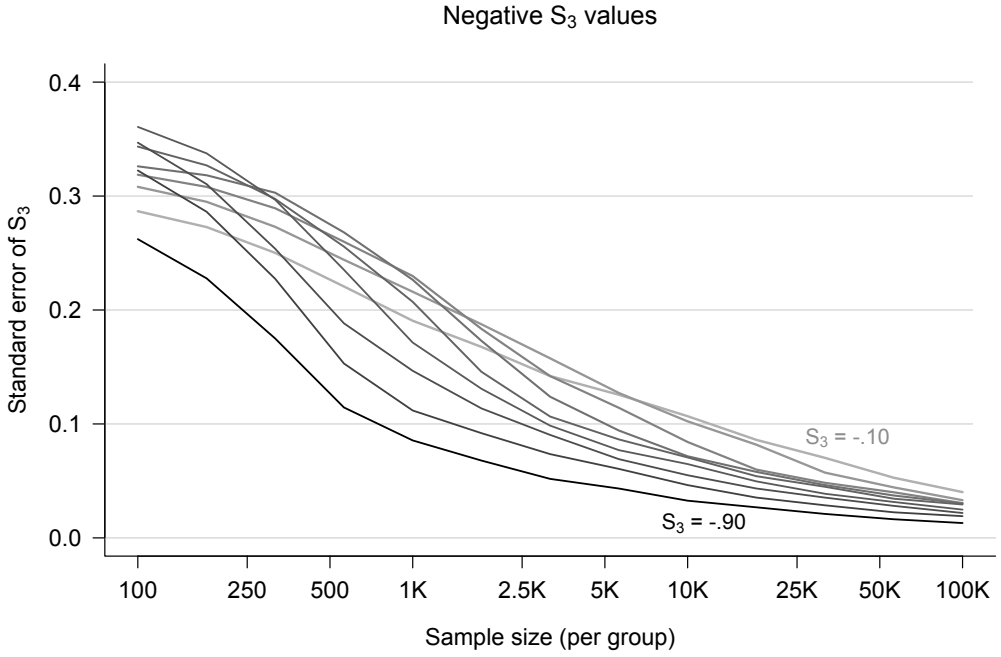

Fig. S7. Standard error estimates for negative values of the population S_3, based on equal-sized samples from two normally distributed populations.