Reaction times and other skewed distributions: problems with the mean and the median

Guillaume A. Rousselet[1] **and Rand R. Wilcox**[2]

[1]Institute of Neuroscience and Psychology, College of Medical, Veterinary and Life Sciences, University of Glasgow, 58 Hillhead Street, G12 8QB, Glasgow, UK
[2]Dept. of Psychology, University of Southern California, Los Angeles, CA 90089-1061, USA

ABSTRACT

To summarise skewed (asymmetric) distributions, such as reaction times, typically the mean or the median are used as measures of central tendency. Using the mean might seem surprising, given that it provides a poor measure of central tendency for skewed distributions, whereas the median provides a better indication of the location of the bulk of the observations. However, the sample median is biased: with small sample sizes, it tends to overestimate the population median. This is not the case for the mean. Based on this observation, Miller (1988) concluded that "sample medians must not be used to compare reaction times across experimental conditions when there are unequal numbers of trials in the conditions." Here we replicate and extend Miller (1988), and demonstrate that his conclusion was ill-advised. In particular, we show that the main source of bias is not a difference in sample size but a difference in skewness. We also demonstrate that bias can be corrected using a percentile bootstrap bias correction. More importantly, a careful examination of the sampling distributions reveals that the sample median is not median biased, whereas the mean is median biased, which implies that in a typical experiment, the median provides a better estimate of central tendency than the mean. All the code and data to reproduce the figures and analyses in the article are available online.

Keywords: mean, median, sampling, bias, bootstrap, estimation, skewness

Introduction

Distributions of reaction times (RT) and many other quantities in social and life sciences are skewed (asymmetric). This asymmetry tends to differ among experimental conditions, such that a measure of central tendency and a measure of spread are insufficient to capture how conditions differ. Instead, to understand the potentially rich differences among distributions, it is advised to consider multiple quantiles of the distributions (Doksum, 1974; Pratte et al., 2010; G. A. Rousselet et al., 2017), or to explicitly model the shapes of the distributions (Heathcote et al., 1991; Rouder et al., 2005; Palmer et al., 2011; Matzke et al., 2013). Yet, it is still common practice to summarise RT distributions using a single number, most often the mean: that one value for each participant and each condition can then be entered into a group ANOVA to make statistical inferences. Because of the skewness of reaction times, the mean is however a poor measure of central tendency: skewness shifts the mean away from the bulk of the distribution, an effect that can be amplified by the presence of outliers or a thick right tail. For instance, in Figure 1A, the median better represents the typical observation than the mean because it is closer to the bulky part of the distribution.

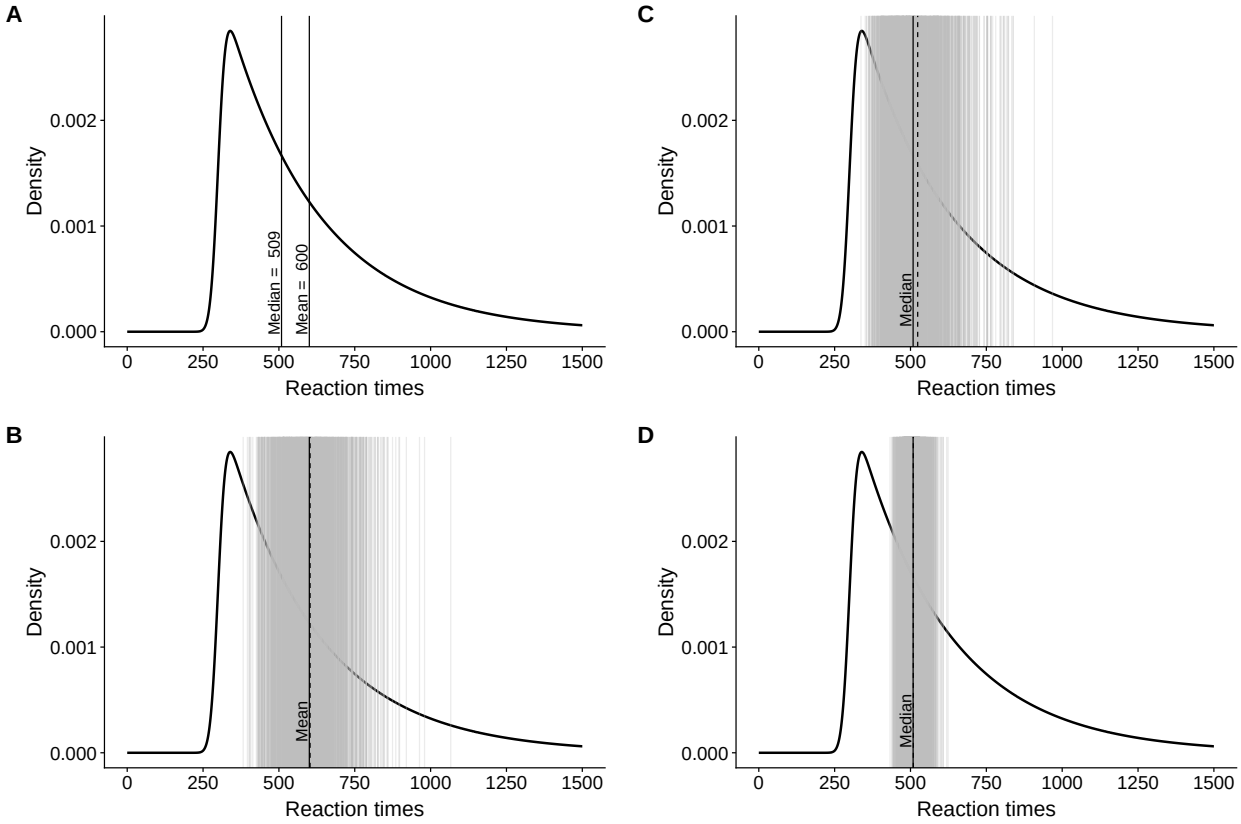

Figure 1. Skewness, sampling and bias. A. Ex-Gaussian distribution with parameters $\mu = 300$, $\sigma = 20$ and $\tau = 300$. The distribution is bounded to the left and has a long right tail. This distribution is used by convenience for illustration, because it looks like a very skewed reaction time distribution. The vertical lines mark the population mean and median. **B.** The vertical grey lines indicate 1,000 means from 1,000 random samples of 10 observations. As in panel A, the vertical black line marks the population mean. The vertical black dashed line marks the mean of the 1,000 sample means. **C.** Same as panel B, but for the median. **D.** Same as C, but for 1,000 samples of 100 observations.

So the median appears to be a better choice than the mean if the goal is to have a single value that reflects the location of most observations in a skewed distribution. The choice between the mean and the median is however more complicated. It could be argued that because the mean is sensitive to skewness, outliers and the thickness of the right tail, it is better able to capture changes in the shapes of the distributions among conditions. But the use of a single value to capture shape differences necessarily leads to intractable analyses because the same mean could correspond to various shapes. If the goal is to understand shape differences between conditions, a multiple quantile approach or explicit shape modelling should be used instead, as mentioned previously.

The mean and the median differ in another important aspect: for small sample sizes, the sample mean is unbiased, whereas the sample median is biased. Concretely, if we perform many times the same RT experiment, and for each experiment we compute the mean and the median, the average mean will be very close to the population mean. As the number of experiments increases, the average sample mean will converge to the exact population mean. This is not the case for the median when sample size is small.

To illustrate, let's imagine that we perform experiments to try to estimate the mean and the median population values of the skewed distribution in Figure 1A. Let's say we take 1,000 samples of 10 observations. For each experiment (sample), we compute the mean. These sample means are shown as grey vertical lines in Figure 1B. A lot of them fall very near the population mean (black vertical line), but some of them are way off. The mean of these estimates is shown with the black dashed vertical line. The difference between the mean of the mean estimates and the population value is called bias. Here bias is small (2.5). Increasing the number of experiments will eventually lead to a bias of zero. In other words, the sample mean is an unbiased estimator of the population mean.

For small sample sizes from skewed distributions, this is not the case for the median. If we proceed as we did for the mean, by taking 1,000 samples of 10 observations, the bias is 15.1: the average median across 1,000 experiments over-estimates the population median (Figure 1C). Increasing sample size to 100 reduces the bias to 0.7 and improves the precision of our estimates. On average, we get closer to the population median, and the distribution of sample medians has much lower variance (Figure 1D).

The reason for the bias of the median is explained by Miller (1988):

> 'Like all sample statistics, sample medians vary randomly (from sample to sample) around the true population median, with more random variation when the sample size is smaller. Because medians are determined by ranks rather than magnitudes of scores, the population percentiles of sample medians vary symmetrically around the desired value of 50%. For example, a sample median is just as likely to be the score at the 40th percentile in the population as the score at the 60th percentile. If the original distribution is positively skewed, this symmetry implies that the distribution of sample medians will also be positively skewed. Specifically, unusually large sample medians (e.g., 60th percentile) will be farther above the population median than unusually small sample medians (e.g., 40th percentile) will be below it. The average of all possible sample medians, then, will be larger than the true median, because sample medians less than the true value will not be small enough to balance out the sample medians greater than the true value. Naturally, the more the distribution is skewed, the greater will be the bias in the sample median.'

Because of this bias, Miller (1988) recommended to not use the median to study skewed distributions in certain situations. As we demonstrate here, this advice was ill-founded. If the choice is between the mean and the median, the median appears to be a better choice for several reasons explored in this article, which is organised in 5 sections. First, we replicate Miller's simulations of estimations from

single distributions. Second, we introduce bias correction and apply it to Miller's simulations. Then we extend Miller's simulations to the actual comparisons of two conditions. Fourth, we examine sampling distributions in detail to reveal unexpected features of the sample mean and median. Finally, we consider a large dataset of RT from a lexical decision task.

1 Replication of Miller 1988

To illustrate the sample median's bias, Miller (1988) employed 12 ex-Gaussian distributions that differed in skewness (Table 1). The distributions are illustrated in Figure 2, and colour-coded using the difference between the mean and the median as a non-parametric measure of skewness. Figure 1 used the most skewed distribution of the 12, with parameters ($\mu = 300$, $\sigma = 20$, $\tau = 300$).

Table 1. Miller's 12 ex-Gaussian distributions. Each distribution is defined by the combination of the three parameters μ (mu), σ (sigma) and τ (tau). The mean is defined as the sum of parameters μ and τ. The median was calculated based on samples of 1,000,000 observations (Miller 1988 used 10,000 observations). Skewness is defined as the difference between the mean and the median.

μ	σ	τ	**mean**	**median**	**skewness**
300	20	300	600	509	92
300	50	300	600	512	88
350	20	250	600	524	76
350	50	250	600	528	72
400	20	200	600	540	60
400	50	200	600	544	55
450	20	150	600	555	45
450	50	150	600	562	38
500	20	100	600	572	29
500	50	100	600	579	21
550	20	50	600	588	12
550	50	50	600	594	6

To estimate bias, following Miller (1988) we performed a simulation in which we sampled with replacement 10,000 times from each of the 12 distributions. We took random samples of sizes 4, 6, 8, 10, 15, 20, 25, 35, 50 and 100, as did Miller. For each random sample, we computed the mean and the median. For each sample size and ex-Gaussian parameter, the bias was then defined as the difference between the mean of the 10,000 sample estimates and the population value.

First, as shown in Figure 3, we can check that the mean is not biased. Each line shows the results for one type of ex-Gaussian distribution: the mean of 10,000 simulations for different sample sizes minus the population mean (600). Irrespective of skewness and sample size, bias is very near zero. The shaded areas mark the upper parts of the 50% highest-density intervals (HDI) of the 10,000 simulations for the least skewed distribution (lighter grey) and the most skewed distribution (darker grey). A HDI is the shortest interval that contains a certain percentage of observations from a distribution (Kruschke, 2013). For symmetric distributions, HDI and confidence intervals are similar, but for skewed distributions, HDI better capture the location of the bulk of the observations. Here, these intervals show the variability across simulations and highlights an important aspect of the results: bias is a long-run property of an estimator; there is no guarantee that one value from a single experiment will be close to the population

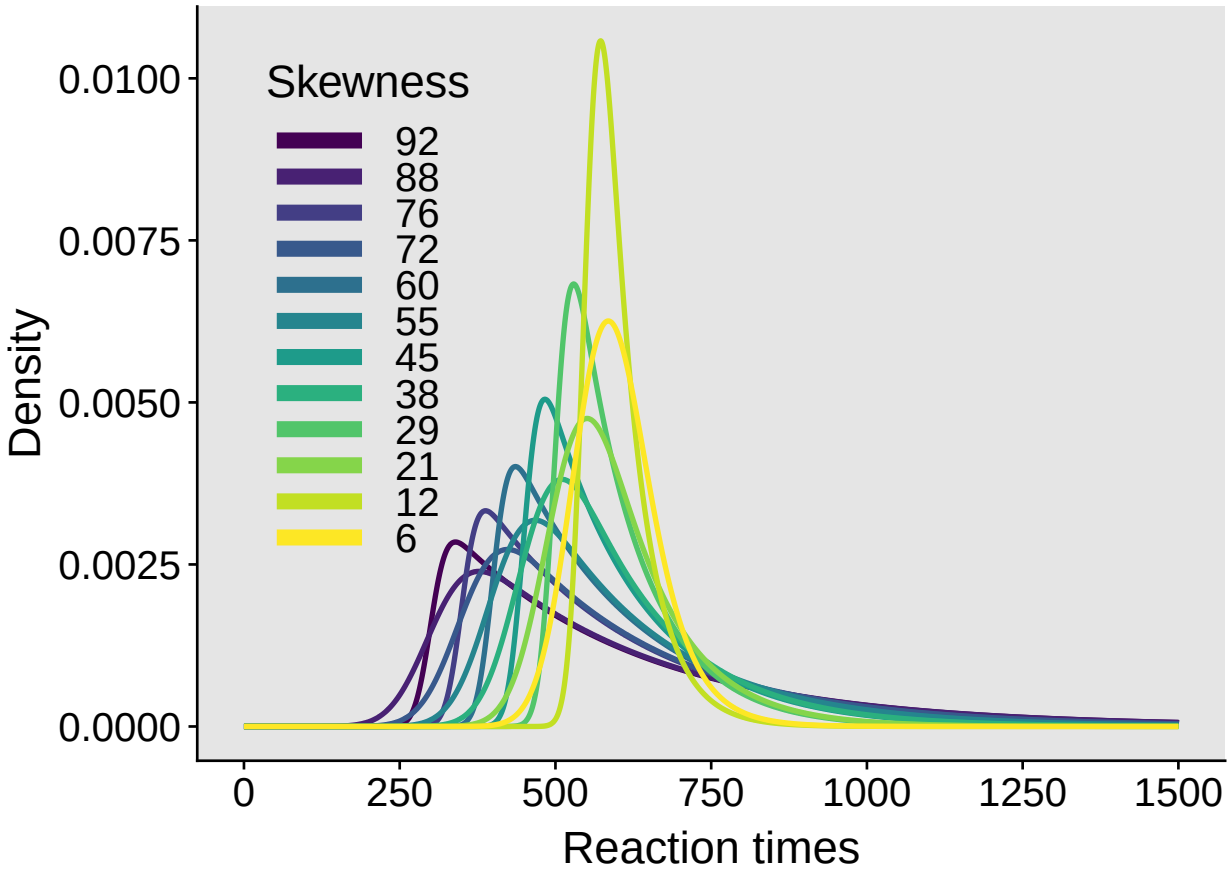

Figure 2. Miller's 12 ex-Gaussian distributions.

value, particularly for small sample sizes. Also, the variability among samples increases with decreasing sample size, which is why results across small n experiments can differ substantially. (We will return to the shape of the sampling distributions in a later section.)

Contrary to the mean, the median estimates are biased for small sample sizes. The values from our simulations are very close to the values reported in Miller (1988) (Table 2).

The results are also illustrated in Figure 3B. As reported by Miller (1988), bias can be quite large and it gets worse with decreasing sample sizes and increasing skewness. Based on these results, Miller (1988) made this recommendation:

> 'An important practical consequence of the bias in median reaction time is that sample medians must not be used to compare reaction times across experimental conditions when there are unequal numbers of trials in the conditions.'

According to Google Scholar, Miller (1988) has been cited 172 times. A look at some of the oldest and most recent citations reveals that his advice has been followed. A popular review article on reaction times, cited 370 times, reiterates Miller's recommendations (Whelan, 2008).

However, there are several problems with Miller's advice, which we explore in the next sections, starting with one key omission from Miller's assessment: the bias of the sample median can be corrected using a percentile bootstrap bias correction.

Table 2. Bias estimation for Miller's 12 ex-Gaussian distributions. Columns correspond to different sample sizes. Rows correspond to different distributions, sorted by skewness. Skewness is defined as the difference between the mean and the median.

Skewness	n=4	n=6	n=8	n=10	n=15	n=20	n=25	n=35	n=50	n=100
92	41	26	19	18	8	8	6	4	3	1
88	39	27	21	16	10	7	5	5	3	2
76	35	23	16	12	8	7	5	4	3	1
72	35	24	16	14	8	6	5	4	3	2
60	28	18	15	9	6	6	4	3	2	1
55	26	18	12	9	7	5	4	3	2	1
45	21	14	10	9	5	4	3	2	2	1
38	18	11	8	7	5	3	3	1	1	1
29	13	10	6	5	3	2	2	1	1	0
21	9	6	4	4	2	2	1	1	1	0
12	5	4	3	2	1	1	1	1	0	0
6	2	2	1	0	1	0	0	0	0	0

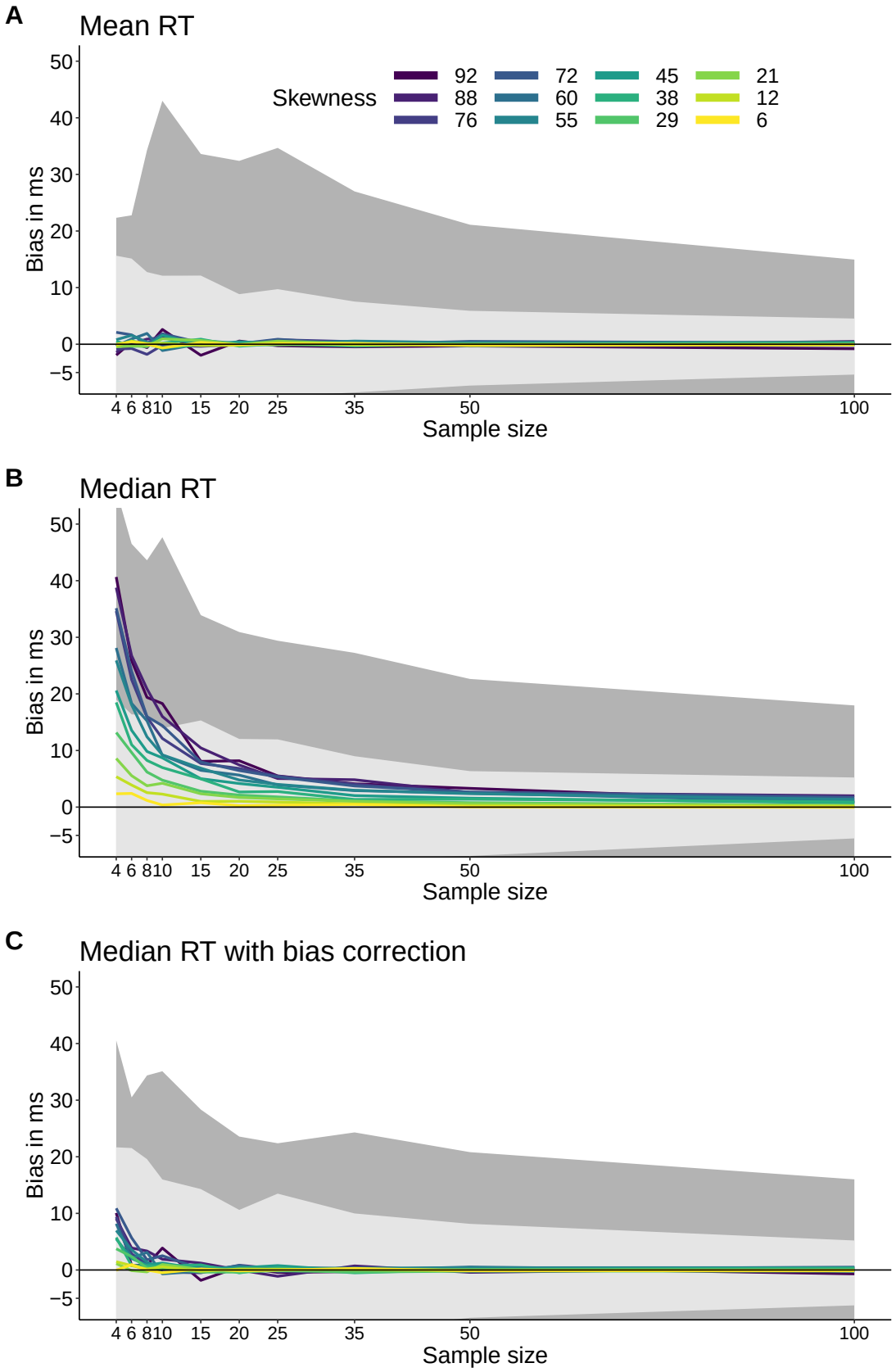

Figure 3. Bias estimation. In each panel, the shaded areas indicate the upper parts of the 50% HDI of the sampling distributions for the least skewed distribution (lighter grey) and the most skewed distribution (darker grey). **A.** Bias for mean reaction times. **B.** Bias for median reaction times. **C.** Bias for median reaction times after bootstrap bias correction.

2 Bias correction

A simple technique to estimate and correct sampling bias is the percentile bootstrap (Efron & Tibshirani, 1994). If we have a sample of n observations, here is how it works:

- sample with replacement n observations from the original sample
- compute the estimate (say the mean or the median)
- perform steps 1 and 2 *nboot* times
- compute the mean of the *nboot* bootstrap estimates

The difference between the estimate computed using the original sample and the mean of the *nboot* bootstrap estimates is a bootstrap estimate of bias.

To illustrate, let's consider one random sample of 10 observations from the skewed distribution in Figure 1A, which has a population median of 508.7 (rounded to 509 in the figure):

$$sample = [355.0, 350.0, 466.7, 1758.2, 604.5, 1707.6, 367.2, 1741.3, 331.4, 1193.2]$$

The median of the sample is 535.6, which over-estimates the population value of 508.7. Next, we sample with replacement 1,000 times from our sample, and compute 1,000 bootstrap estimates of the median. The distribution obtained is a bootstrap estimate of the sampling distribution of the median, as illustrated in Figure 4. The idea is this: if the bootstrap distribution approximates, on average, the shape of the sampling distribution of the median, then we can use the bootstrap distribution to estimate the bias and correct our sample estimate. However, as we're going to see, this works on average, in the long run. There is no guarantee for a single experiment.

The mean of the bootstrap estimates is 722.6. Therefore, our estimate of bias is the difference between the mean of the bootstrap estimates and the sample median, which is 187, as shown by the black horizontal arrow in Figure 4. To correct for bias, we subtract the bootstrap bias estimate from the sample estimate (grey horizontal arrow in Figure 4):

sample median - (mean of bootstrap estimates - sample median)

which is the same as:

2 x sample median - mean of bootstrap estimates.

Here the bias corrected sample median is 348.6. So the sample bias has been reduced dramatically, clearly too much from the original 535.6. But bias is a long run property of an estimator, so let's look at a few more examples. We take 100 samples of n = 10, and compute a bias correction for each of them. The results of these 100 virtual experiments are shown in Figure 5. The arrows go from the sample median to the bias corrected sample median. The black vertical line shows the population median we're trying to estimate.

With n = 10, the sample estimates have large spread around the population and more so on the right than the left of the distribution. The bias correction also varies a lot in magnitude and direction, sometimes improving the estimates, sometimes making matters worse. Across experiments, it seems that the bias correction was too strong: the population median was 508.7, the average sample median was 515.1, but the average bias corrected sample median was only 498.8.

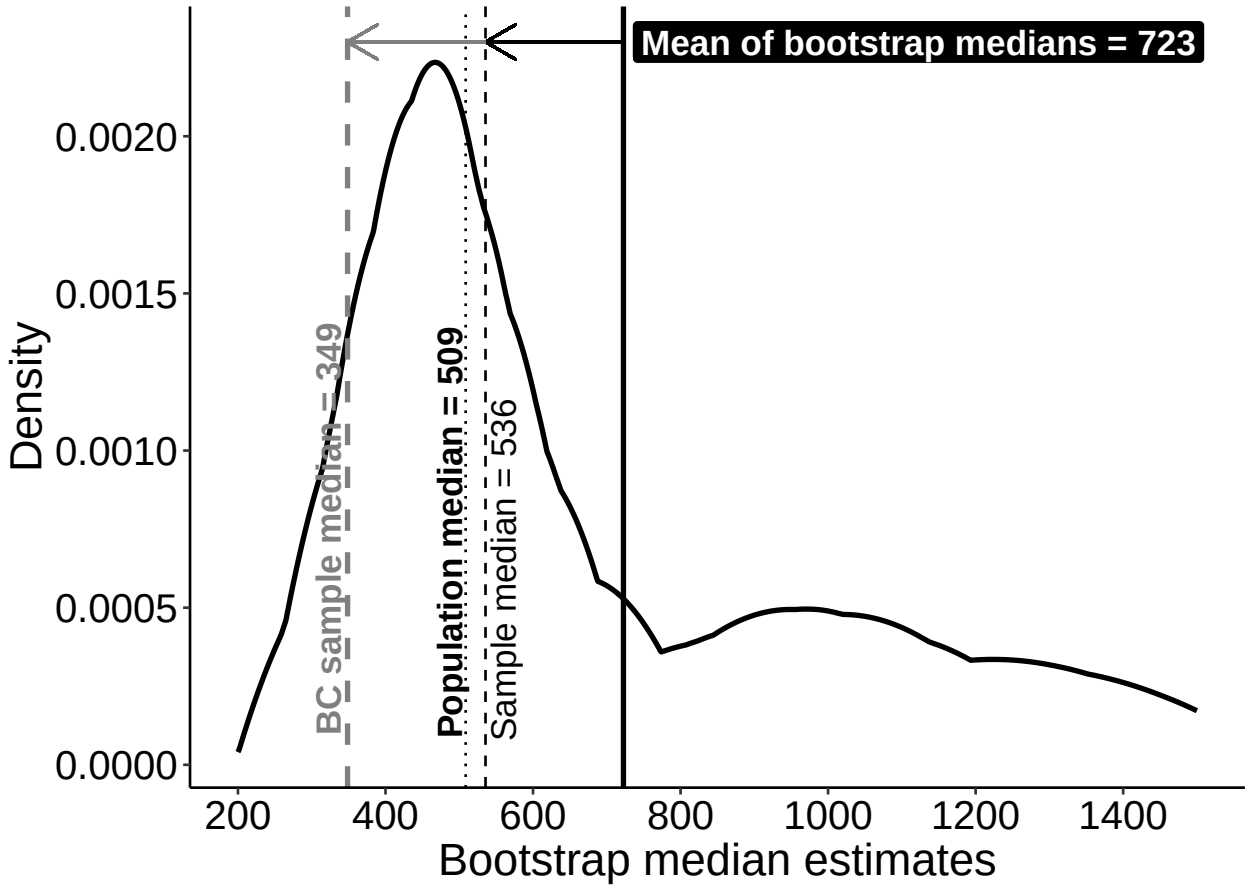

Figure 4. Bootstrap bias correction example: one experiment. The sample median (black dashed vertical line) overestimates the population value (black dotted vertical line). The kernel density estimate of 1,000 bootstrap estimates of the sample median suggests, correctly, that the median sampling distribution is positively skewed. The difference between the sample median and the mean of the bootstrap medians (thick black vertical line) defines the bootstrap estimate of the bias (black horizontal arrow). This estimate can be subtracted from the sample median (grey horizontal arrow) to obtain a bias corrected (BC) sample median (grey dashed vertical line).

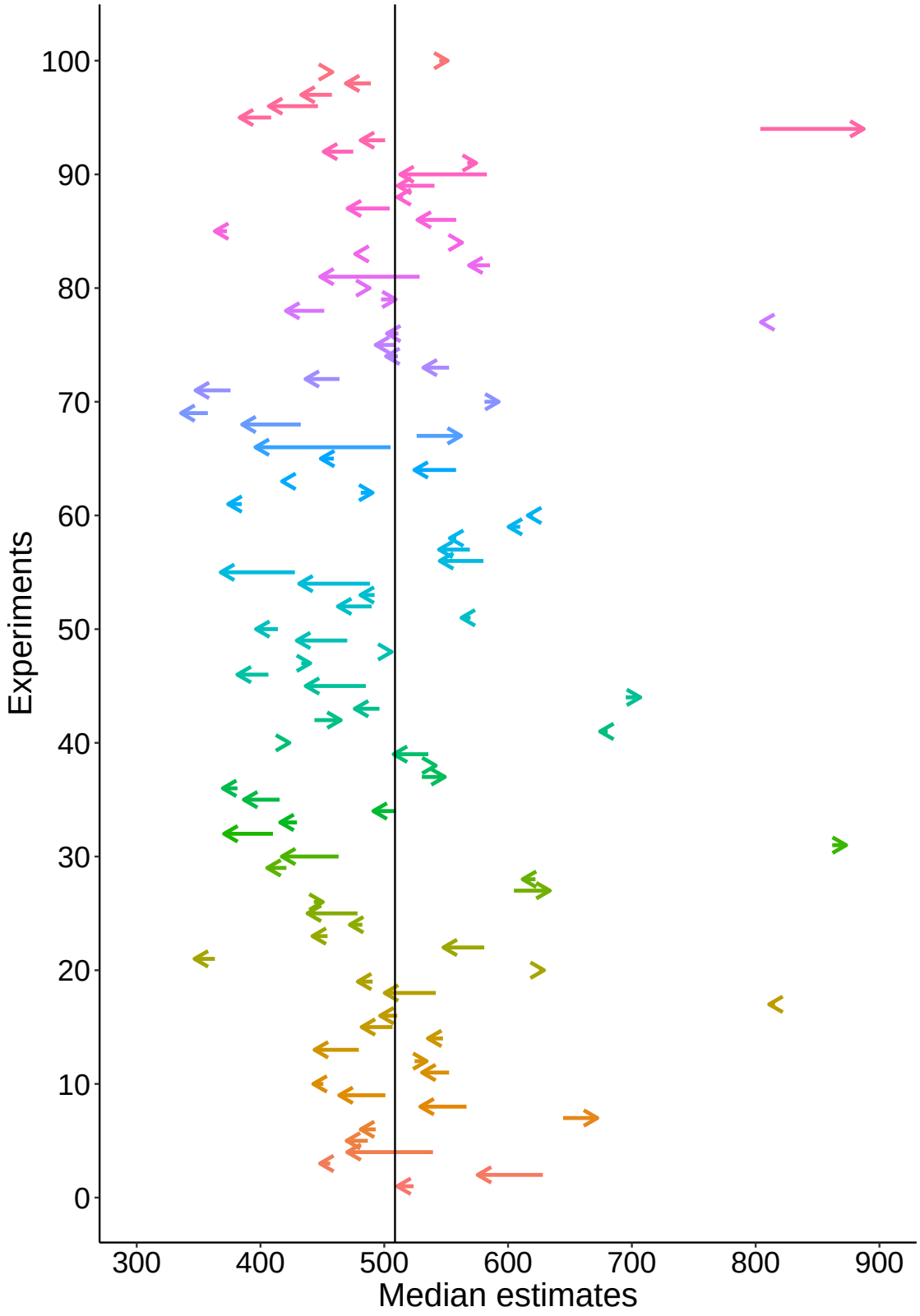

Figure 5. Bootstrap bias correction example: 100 experiments For each experiment, 10 observations were sampled. Each arrow starts at the sample median for one experiment and ends at the bias corrected sample median. The bias was estimated using 200 bootstrap samples. The black vertical line marks the population median.

What happens if instead of 100 experiments, we perform 1000 experiments, each with n=10, and compute a bias correction for each one? Now the average of the bias corrected median estimates is much closer to the true median: the population median was 508.7, the average sample median was 522.1, and the average bias corrected sample median was 508.6. So the bias correction works very well in the long run for the median. But that's not always the case: it depends on the estimator and on the amount of skewness.

If we apply the bias correction technique to our median estimates of samples from Miller's 12 distributions, we get the results in Figure 3C. For each iteration in the simulation, bias correction was performed using 200 bootstrap samples. The bias correction works very well on average, except for the smallest sample sizes. The failure of the bias correction for very small n is not surprising, because the shape of the sampling distribution cannot be properly estimated by the bootstrap from so few observations. From n = 10, the bias values are very close to those observed for the mean. So it seems that in the long-run, we can eliminate the bias of the sample median by using a simple bootstrap procedure. As we will see in the next section, the bootstrap bias correction is also effective when comparing two groups.

3 Extension of Miller 1988

The sample median is biased when sampling from skewed distributions. The bias increases with decreasing sample size. According to Miller (1988), because of this bias, group comparison can be affected if the two groups differ in sample size, such that real differences can be lowered or increased, and non-existent differences suggested. As a result, for unequal n, Miller advised against the use of the median. In this section, we will see that this advice is wrong for two reasons.

We assessed the problem using a simulation in which we drew samples of same or different sizes from populations that differed in skewness or not. We used the same 12 distributions used by Miller (1988), as described previously. Group 2 had a constant size of n = 200 and was sampled from the distribution with the least skewness. Group 1 had size n = 10 to n = 200, in increments of 10, and was sampled from the 12 distributions. For the mean, the results of 10,000 iterations are presented in Figure 6. All the bias values are near zero, as expected. The lighter grey shaded area shows the upper part of the mean's bias 50% HDI, when group 1 and group 2 have the same, least, skewness (6). The darker grey area shows the upper part of the HDI in the case where group 1 was sampled from the most skewed distribution (92). These intervals show the location of the bulk of the 10,000 simulations. Again, this is a reminder that bias is defined in the long run: a single experiment (one of our simulation iterations) can be far off the population value, especially with small sample sizes.

Results for the median are presented in Figure 6B. Bias increases with skewness and sample size difference (the difference gets larger as the sample size of group 1 gets smaller). However, if the two groups have the same skewness (skewness 6), there is almost no bias even when group 2 has 200 observations and group 1 has only 10 observations. So it seems Miller's (1988) warning about differences in sample sizes was inappropriate, because the main factor causing bias is skewness.

Next, let's find out if we can correct the bias. Bias correction was performed in 2 ways: with the bootstrap, as explained in the previous section, and with a different approach using subsamples. The second approach was suggested by Miller (1988):

> 'Although it is computationally quite tedious, there is a way to use medians to reduce the effects of outliers without introducing a bias dependent on sample size. One uses the regular median from Condition F and compares it with a special "average median" (Am) from Condition M. To compute Am, one would take from Condition M all the possible subsamples

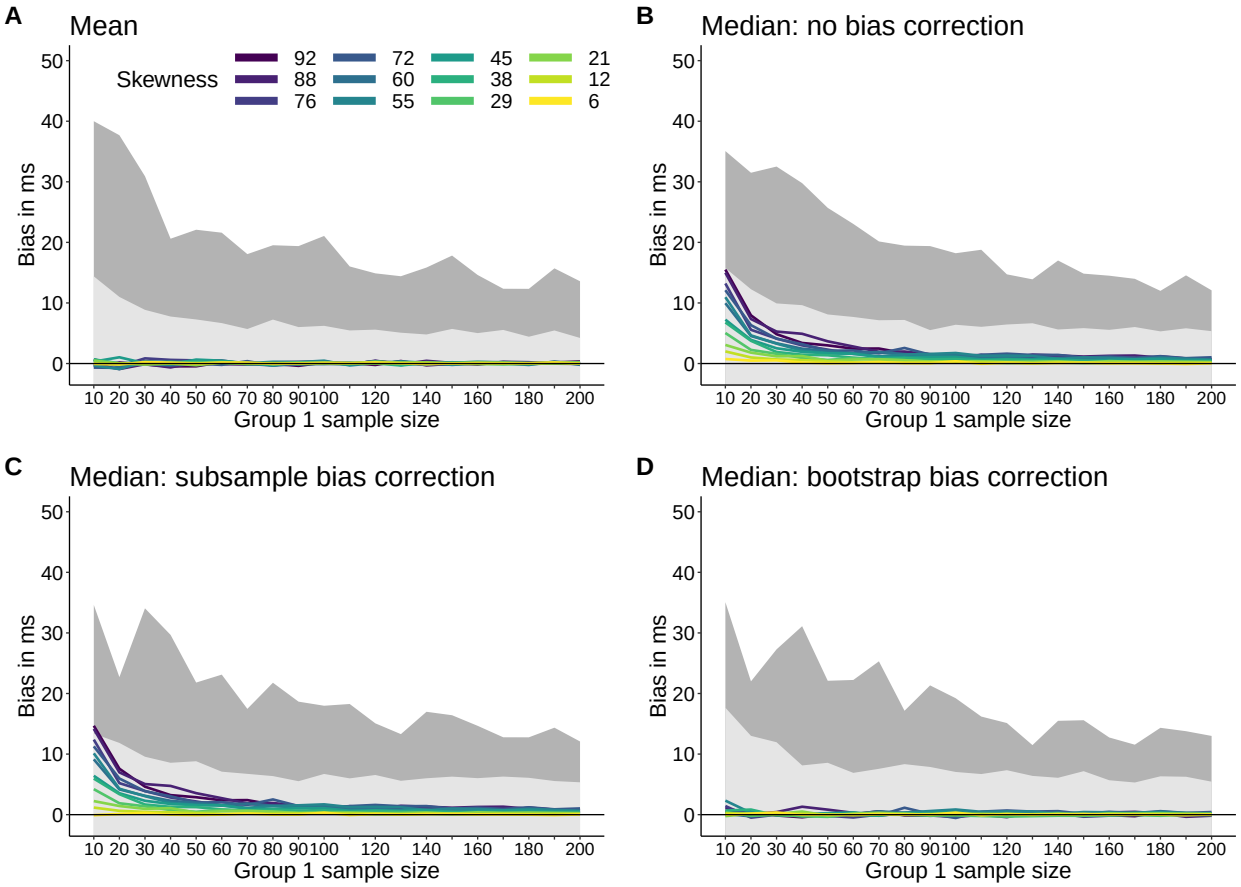

Figure 6. Bias estimation for the difference between two independent groups. Group 2 is always sampled from the least skewed distribution and has size n = 200. The size of group 1 is indicated along the x axis in each panel. Group 1 is sampled from Miller's 12 skewed distributions and the results are colour coded by skewness. The shaded areas indicate the 50% HDIs of the sampling distributions for the least skewed distribution (lighter grey) and the most skewed distribution (darker grey). **A.** Bias for mean reaction times. **B.** Bias for median reaction times. **C.** Bias for median reaction times after subsample bias correction. **D.** Bias for median reaction times after bootstrap bias correction.

> of Size f where f is the number of trials in Condition F. For each subsample one computes the subsample median. Then, Am is the average, across all possible subsamples, of the subsample medians. This procedure does not introduce bias, because all medians are computed on the basis of the same sample (subsample) size.'

Using all possible subsamples would take far too long; for instance, if one group has 5 observations and the other group has 20 observations, there are 15504 subsamples to consider (*choose(20,5)* in R). Slightly larger sample sizes would force us to consider millions of subsamples. So instead we computed K random subsamples. We arbitrarily set K to 1,000. Although this is not what Miller (1988) suggested, the K loop shortcut should reduce bias to some extent if it is due to sample size differences. The results are presented in Figure 6C. The K loop approach has only a very limited effect on bias. The reason is simple: the main cause of the bias is not a difference in sample size, it is a difference in skewness.

The skewness difference can be handled by the bootstrap. Bias correction using 200 bootstrap samples for each simulation iteration leads to the results in Figure 6D: overall the bootstrap bias correction works very well. For instance, for the most skewed distribution and n = 10, the mean's bias is -0.51, whereas the median's bias after bias correction is 0.49. None of them are exactly zero and the absolute values are very similar. At most, for n = 10, the median's maximum bias across distributions is 2.32 ms, whereas the mean's is 0.7 ms.

In conclusion, Miller's (1988) advice was inappropriate because, when comparing two groups, bias originates from a conjunction of differences in skewness and in sample size, not from differences in sample size alone. Also, the bias can be corrected using the bootstrap. In practice, we would expect that skewed distributions from different conditions/groups differ in skewness, in which case unequal sample sizes will exacerbate the bias. To be cautious, when sample size is relatively small, it could be useful to report median effects with and without bootstrap bias correction. It would be even better to run simulations to determine the sample sizes required to achieve an acceptable bias value.

4 Sampling distributions

Bias is defined as the distance between the mean of the sampling distribution (here estimated using Monte-Carlo simulations) and the population value. In the previous sections, we saw that for small sample sizes, the sample median provides a biased estimation of the population median, which can significantly affect group comparisons. However, this bias disappears with large sample sizes, and it can be corrected using a bootstrap bias correction. In this section, we look in more detail at the shape of the sampling distributions, which was ignored by Miller (1988).

Let's consider the sampling distributions of the mean and the median for different sample sizes from the ex-Gaussian distributions described in Figure 2. When skewness is low (6, first row of Figure 7), the sampling distributions are symmetric and centred on the population values: there is no bias. As we saw previously, with increasing sample size, variability decreases, which is why studies with larger samples provide more accurate estimations. The flip side is that studies with small samples are much noisier, which is why their results tend to not replicate (Button et al., 2013).

When skewness is large (92, second row of Figure 7), sampling distributions get more positively skewed with decreasing sample sizes. To better understand how the sampling distributions change with sample size, we turn to the last row of Figure 7, which shows 50% HDI. Each horizontal line is a HDI for a particular sample size. The labels contain the values of the interval boundaries. The coloured vertical tick inside the interval marks the median of the distribution. The red vertical line spanning the entire plot is the population value.

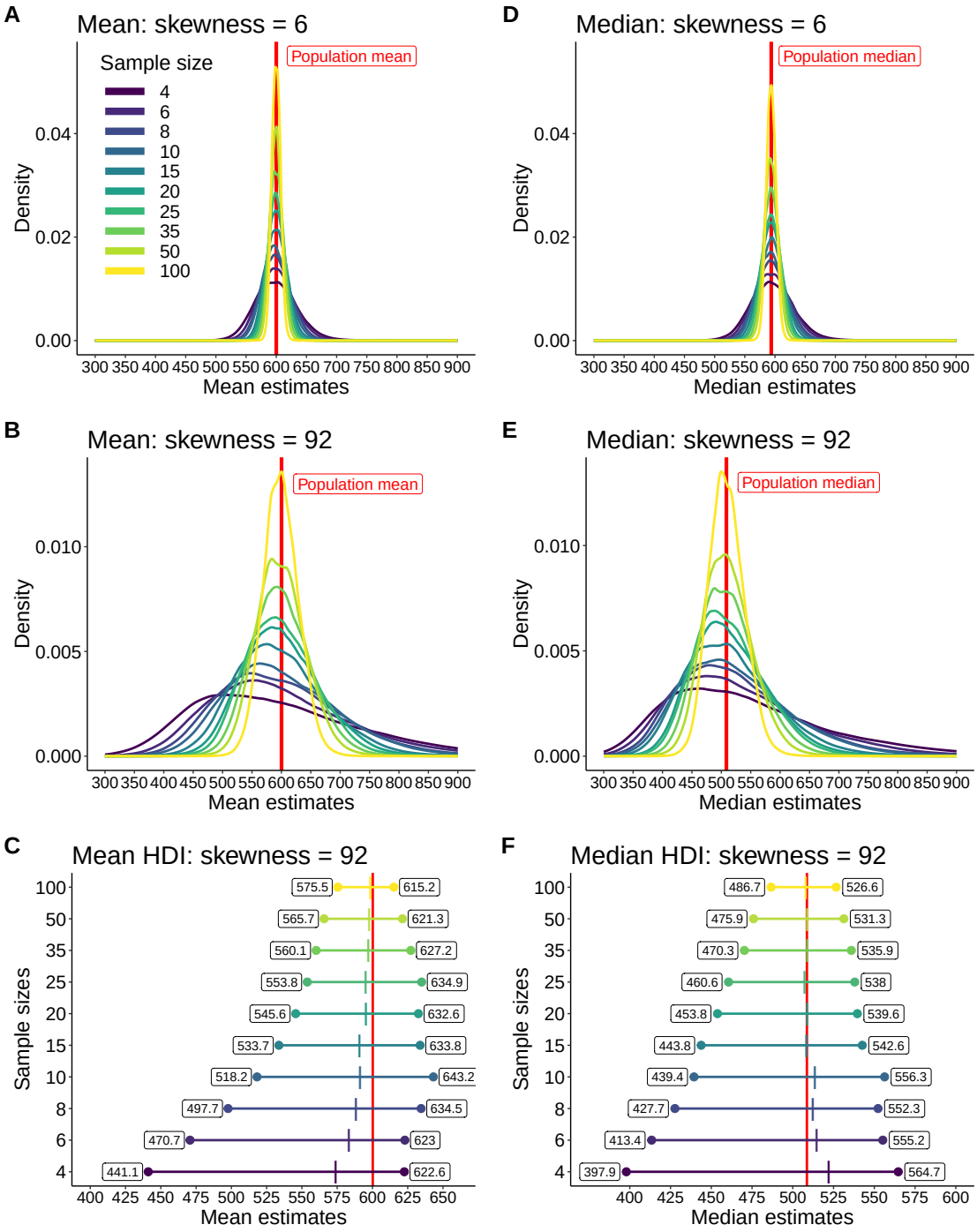

Figure 7. Sampling distributions of the mean and the median. In each panel, results for sample sizes from 4 to 100 are colour coded. The results are based on 10,000 samples for each sample size and skewness. Left column: results for the mean; right column: results for the median. The bottom row illustrates, for the mean (**C**) and the median (**F**), the 50% HDI of the distributions shown in the second row.

For means and small sample sizes, the 50% HDI is offset to the left of the population mean, and so is the median of the sampling distribution. This demonstrates that the typical sample mean tends to under-estimate the population mean – that is to say, the mean sampling distribution is median biased. This offset reduces with increasing sample size, but is still present even for n=100.

For medians and small sample sizes, there is a discrepancy between the 50% HDI, which is shifted to the left of the population median, and the median of the sampling distribution, which is shifted to the right of the population median. This contrasts with the results for the mean, and can be explained by differences in the shapes of the sampling distributions, in particular the larger skewness and kurtosis of the median sampling distribution compared to that of the mean. The offset between the sample median and the population value reduces quickly with increasing sample size. For n=10, the median bias is already very small. From n=15, the median sample distribution is not median bias, which means that the typical sample median is not biased.

Another representation of the sampling distributions is provided in Figure 8: 50% HDI of the biases are shown as a function of sample size. For both the mean and the median, the spread of the bias increases with increasing skewness and decreasing sample size. Skewness also increases the asymmetry of the bias distributions, but more so for the mean than the median.

So is the mean also biased? According to the standard definition of bias, which is based on the distance between the population mean and the average of the sampling distribution of the mean, the mean is not biased. But this definition applies to the long run, after we replicate the same experiment many times - 10,000 times in our simulations. So what happens in practice, when we perform only one experiment instead of 10,000? In that case, the median of the sampling distribution provides a better description of the typical experiment than the mean of the distribution. And the median of the sampling distribution of the mean is inferior to the population mean when sample size is small. So if we conduct one small n experiment and compute the mean of a skewed distribution, we're likely to under-estimate the true value.

Is the median biased after all? The median is indeed biased according to the standard definition. However, with small n, the typical median (represented by the median of the sampling distribution of the median) is close to the population median, and the difference disappears for even relatively small sample sizes.

5 Application to a large dataset

The simulations above suggest that using the median is more appropriate than using the mean when dealing with skewed distributions, contrary to Miller's advice. But what happens when we deal with real RT distributions instead of simulated ones? To find out, in this last section, we look at median bias in a large dataset of reaction times from participants engaged in a lexical decision task. The data are from the French lexicon project (FLP) (Ferrand et al., 2010). After removing a few participants who didn't pay attention to the task (low accuracy or many very late responses), we're left with 959 participants. Each participant had between 996 and 1001 trials for each of two conditions, Word and Non-Word. Figure 9 illustrates reaction time distributions from 100 randomly sampled participants in the Word and Non-Word conditions.

Among participants, the variability in the shapes of the distributions is particularly striking. The shapes of the distributions also differed between the Word and the Non-Word conditions. In particular, skewness tended to be larger in the Word than the Non-Word condition. Based on the standard parametric definition of skewness, that was the case in 80% of participants. If we use a non-parametric estimate instead (mean – median), it was the case in 70% of participants. This difference in skewness between conditions implies that the difference between medians will be biased in individual participants.

If we save the median response time for each participant and each condition, we get two distributions

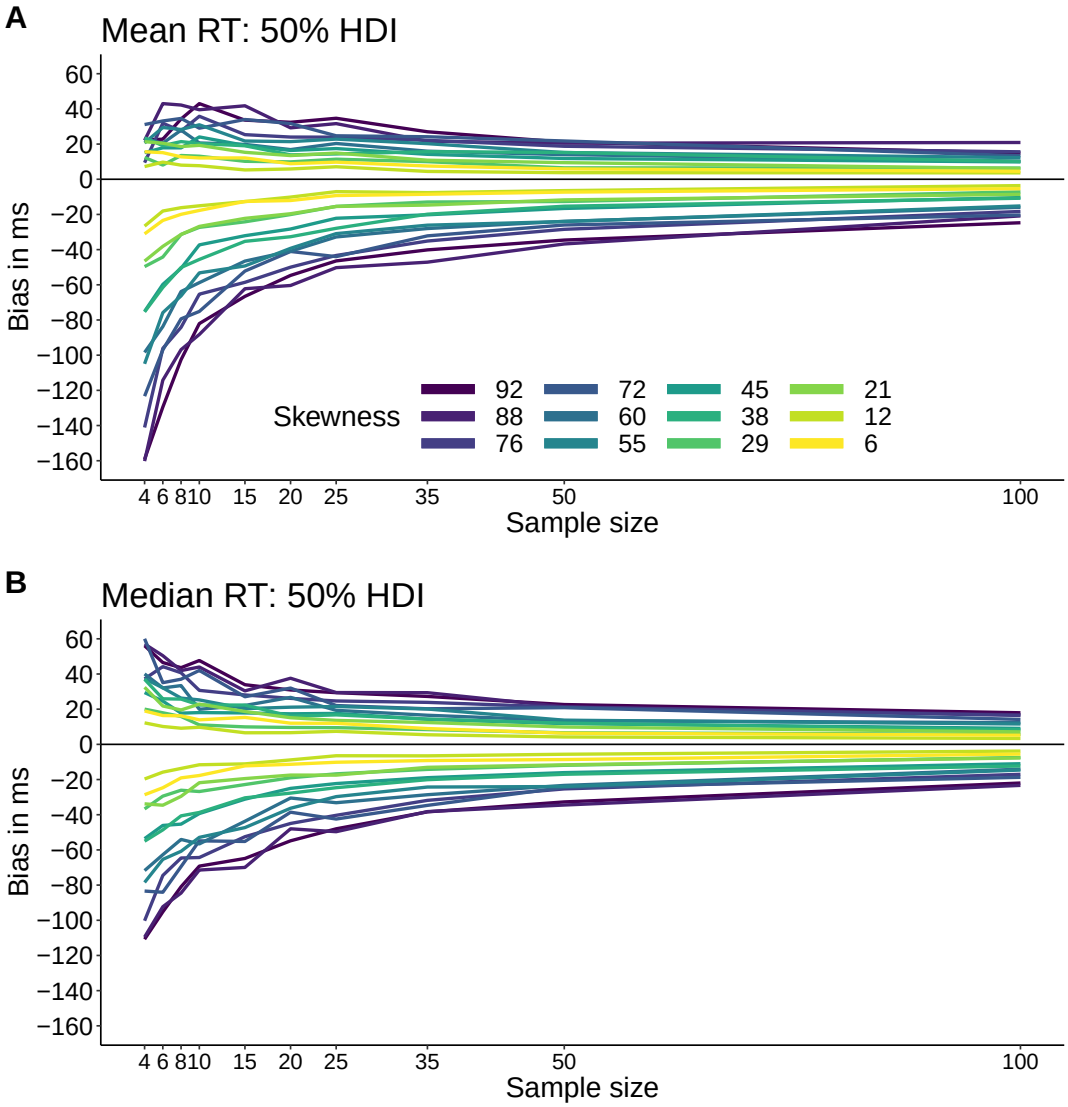

Figure 8. 50% highest density intervals of the biases of the sample mean and the sample median as a function of sample size and skewness.

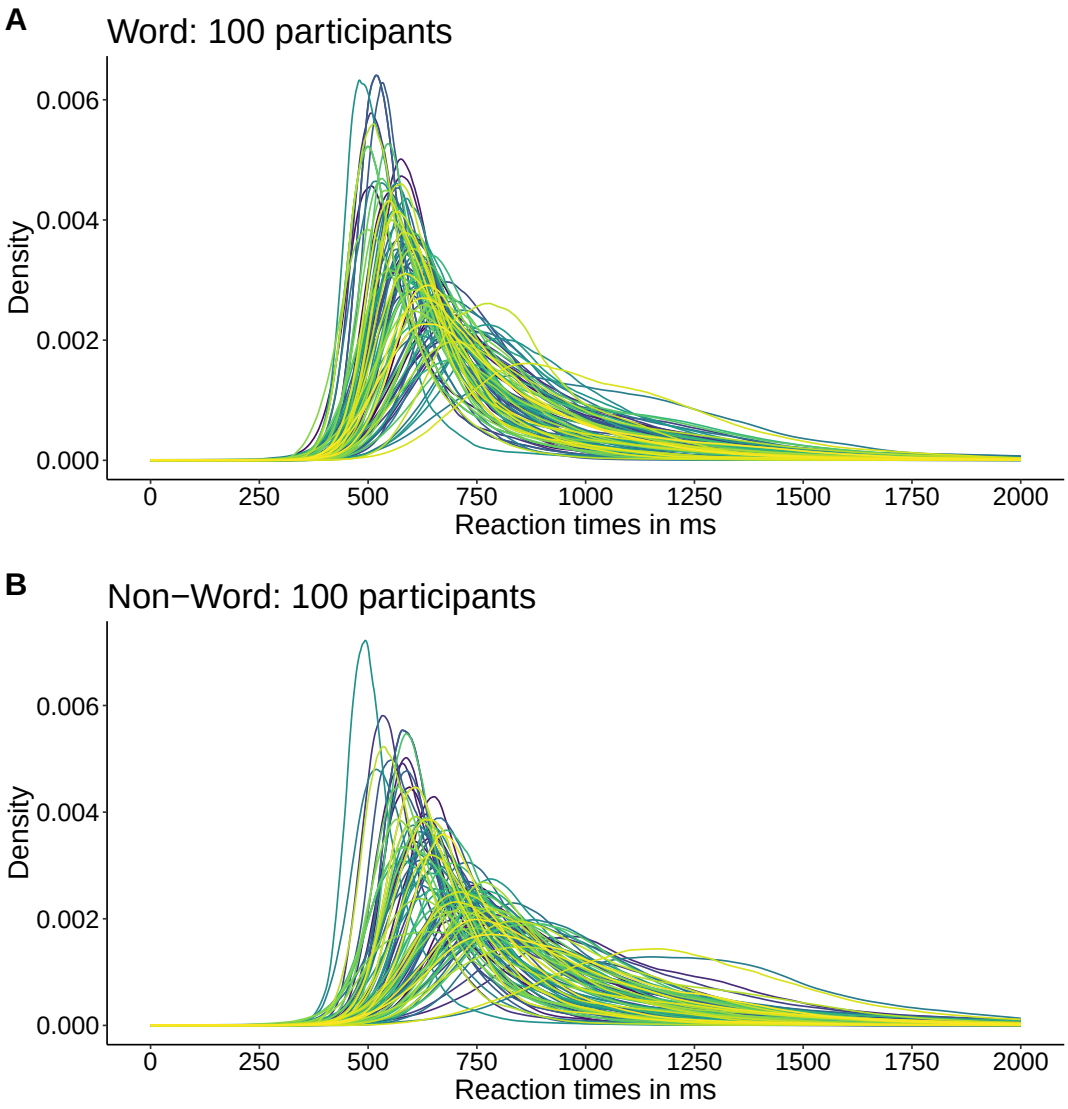

Figure 9. FLP dataset: reaction time distributions from 100 participants. Participants were randomly selected among 959. Distributions are shown for the same participants (colour coded) in the Word (**A**) and Non-Word (**B**) conditions.

that display positive skewness (Figure 10A). The same applies to distributions of means (Figure 10B). The distributions of pairwise differences between the Non-Word and Word conditions is also positively skewed (Figure 10C).

Hence, Figure 9 and Figure 10 demonstrates that we have to worry about skewness at 2 levels of analysis: in individual distributions and in group distributions. Here we explore estimation bias as a result of skewness and sample size in individual distributions, because it is the most similar to Miller's simulations. From what we've learnt so far, we can already make predictions: because skewness tended to be stronger in the Word than in the Non-Word condition, the bias of the median will be stronger in the former than the later for small sample sizes. That is, the median in the Word condition will tend to be more over-estimated than the median in the Non-Word condition. As a consequence, the difference between the median of the Non-Word condition (larger RT) and the median of the Word condition (smaller RT) will tend to be under-estimated. To check this prediction, we estimated bias in every participant using a simulation with 2,000 iterations. We used the full sample of roughly 1,000 trials as the population, from which we computed population means and population medians. Because the Non-Word condition is the least skewed, we used it as the reference condition, which always had 200 trials. The Word condition had 10 to 200 trials, with 10 trial increments. In the simulation, single RT were sampled with replacements among the roughly 1,000 trials available per condition and participant, so that each iteration is equivalent to a fake experiment.

Let's look at the results for the median. Figure 11A shows the bias of the difference between medians (Non-Word – Word), as a function of sample size in the Word condition. The Non-Word condition always had 200 trials. All participants are superimposed and shown as coloured traces. The average across participants is shown as a thicker black line.

As expected, bias tended to be negative with small sample sizes; that is, the difference between Non-Word and Word was underestimated because the median of the Word condition was overestimated. For the smallest sample size, the average bias was -10.9 ms. That's probably substantial enough to seriously distort estimation in some experiments. Also, variability is high, with a 80% highest density interval of [-17.1, -2.6] ms. Bias decreases rapidly with increasing sample size. For n=20, the average bias was -4.8 ms, for n=60 it was only -1 ms.

After bootstrap bias correction (with 200 bootstrap samples), the average bias dropped to roughly zero for all sample sizes (Figure 11B). Bias correction also reduced inter-participant variability.

As we saw in Figure 7, the sampling distribution of the median is skewed, so the standard measure of bias (taking the mean across simulation iterations) does not provide a good indication of the bias we can expect in a typical experiment. If instead of the mean, we compute the median bias, we get the results in Figure 11C. At the smallest sample size, the average bias is only -1.9 ms, and it dropped to -0.3 for n=20. This result is consistent with the simulations reported above and confirms that in the typical experiment, the bias associated with the median is negligible.

What happens with the mean? The average bias of the mean is near zero for all sample sizes (Figure 11D). As we did for the median, we also considered the median bias of the mean (Figure 11E). For the smallest sample size, the average bias across participants is 6.9 ms. This positive bias can be explained from the results using the ExGaussian distributions: because of the larger skewness in the Word condition, the sampling distribution of the mean was more positively skewed for small samples in that condition compared to the Non-Word condition, with the bulk of the bias estimates being negative. That is, the mean tended to be more under-estimated in the Word condition, leading to larger Non-Word – Word differences in the typical experiment.

The results from the real RT distributions confirm our earlier simulations using exGaussian distributions: for small sample sizes, the mean and the median differ in bias due to differences in sampling distributions

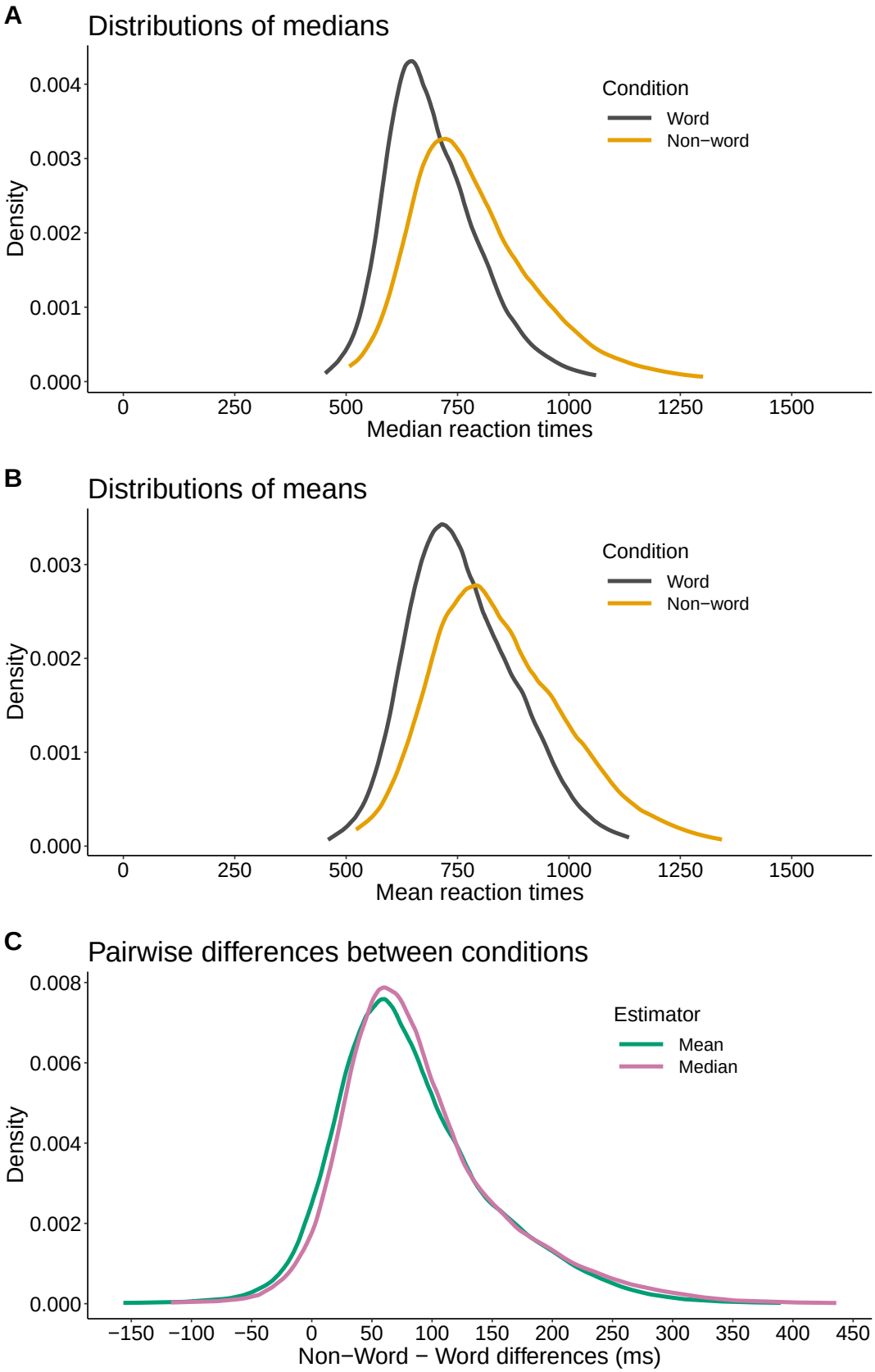

Figure 10. FLP dataset: group distributions. For every participant, the median (**A**) and the mean (**B**) were computed for the Word and Non-Word observations separately. Panel (**C.**) Distributions of pairwise differences between the Non-Word and Word conditions.

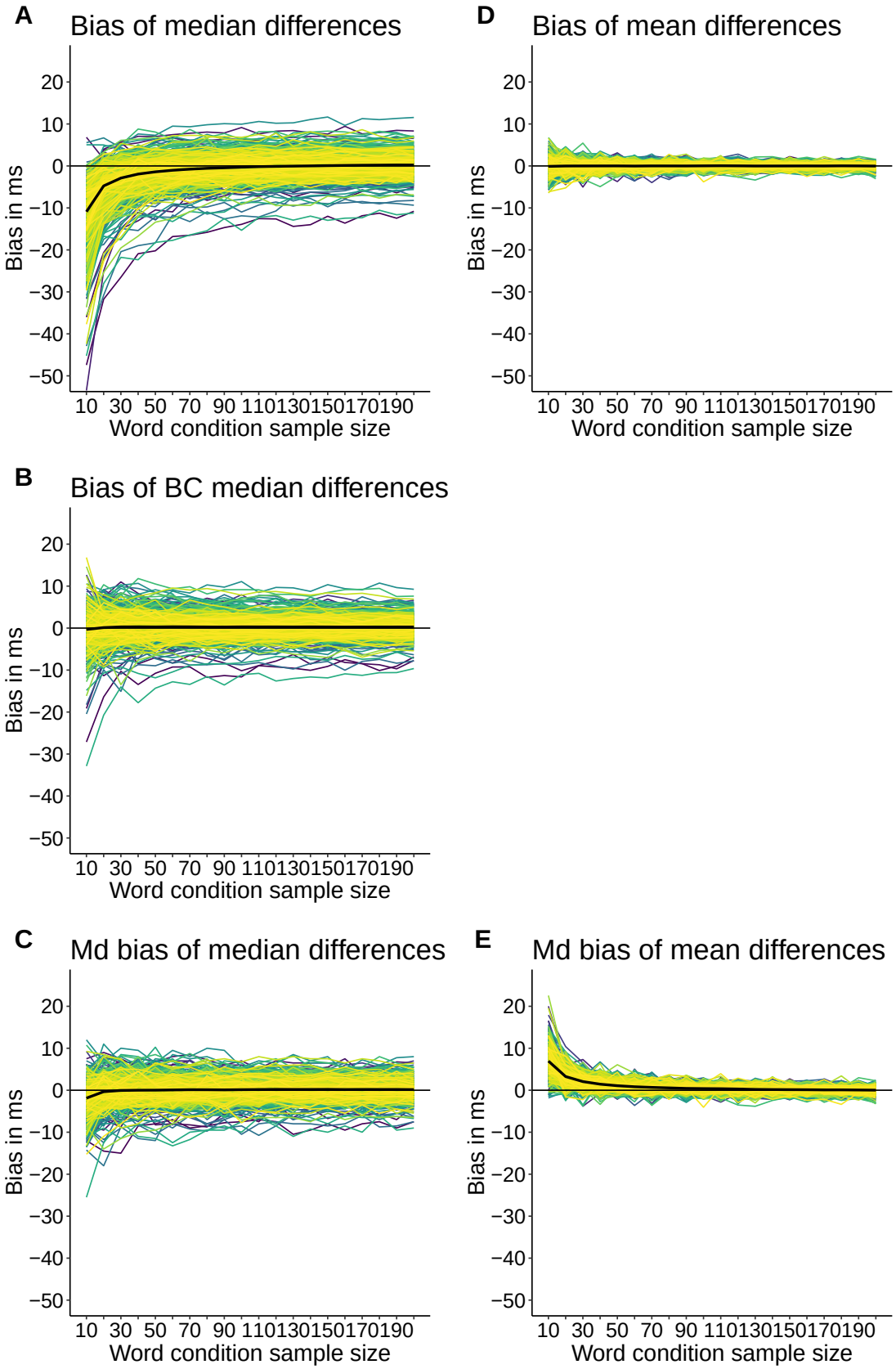

Figure 11. FLP dataset: bias estimation for the difference between the Non-Word and Word conditions. In each panel, thin coloured lines indicate results from individual participants and the thick black line indicates the mean across participants. The left column illustrates results for the median, the right column for the mean. BC = bias-corrected. MD bias = median bias.

and the bias of the median can be corrected using the bootstrap. Another striking difference between the mean and the median is the spread of bias values across participants, which is much larger for the median than the mean. This difference in bias variability does not reflect a difference in variability among participants for the two estimators of central tendency. Indeed, as we saw in Figure 10C, the distributions of differences between Non-Word and Word conditions are very similar for the mean and the median. Estimates of spread are also similar between difference distributions (median absolute deviation to the median (MAD): mean RT = 57 ms; median RT = 54 ms). This suggests that the inter-participant bias differences are due to the individual differences in shape distributions observed in Figure 9, to which the mean and the median are differently sensitive.

The larger inter-participant variability in bias for the median compared to the mean could also suggest that across participants, measurement precision would be lower for the median. We directly assessed measurement precision at the group level by performing a multi-level simulation. In this simulation, we asked, for instance, how often the group estimate was no more than 10 ms from the population value across many experiments (here 10,000). In each iteration (fake experiment) of the simulation, there were 200 trials per condition and participant, such that bias at the participant level was not an issue (a total of 400 trials for an RT experiment is perfectly reasonable and could be done in no more than 20 minutes per participant). For each participant and condition, the mean and the median were computed across the 200 random trials for each condition, and then the Non-Word - Word difference was saved. Group estimation of the difference was based on a random sample of 10 to 300 participants, with the group mean computed across participants' differences between means and the group median computed across participants' differences between medians. Population values were defined by first computing, for each condition, the mean and the median across all available trials for each participant, second by computing across all participants the mean and the median of the pairwise differences. Measurement precision was calculated as the proportion of experiments in which the group estimate was no more than x ms from the population value, with x varying from 5 to 40 ms.

Not surprisingly, the proportion of estimates close to the population value increases with the number of participants for the mean and the median (Figure 12). More interestingly, the relationship was non-linear, such that a larger gain in precision was achieved by increasing sample size for instance from 10 to 20 compared to from 90 to 100. The results also let us answer useful questions for planning experiments (see the black arrows in Figure 12A & B):

- So that in 70% of experiments the group estimate is no more than 10 ms from the population value, we need to test at least 59 participants for the mean, 56 participants for the median.
- So that in 90% of experiments the group estimate is no more than 20 ms from the population value, we need to test at least 37 participants for the mean, 38 participants for the median.

Finally, the mean and the median differed very little in measurement precision (Figure 12C), which suggests that at the group level, the mean does not provide any clear advantage over the median. Of course, different results could be obtained in different situations. For instance, the same simulations could be performed using different numbers of trials per participant and condition. Also, skewness can differ much more among conditions in certain tasks, such as in difficult visual search tasks (Palmer et al., 2011).

Discussion

In this article, we reproduced the simulations from Miller (1988), extended them and applied a similar approach to a large dataset of reaction times (Ferrand et al., 2010). The two sets of analyses led to the same

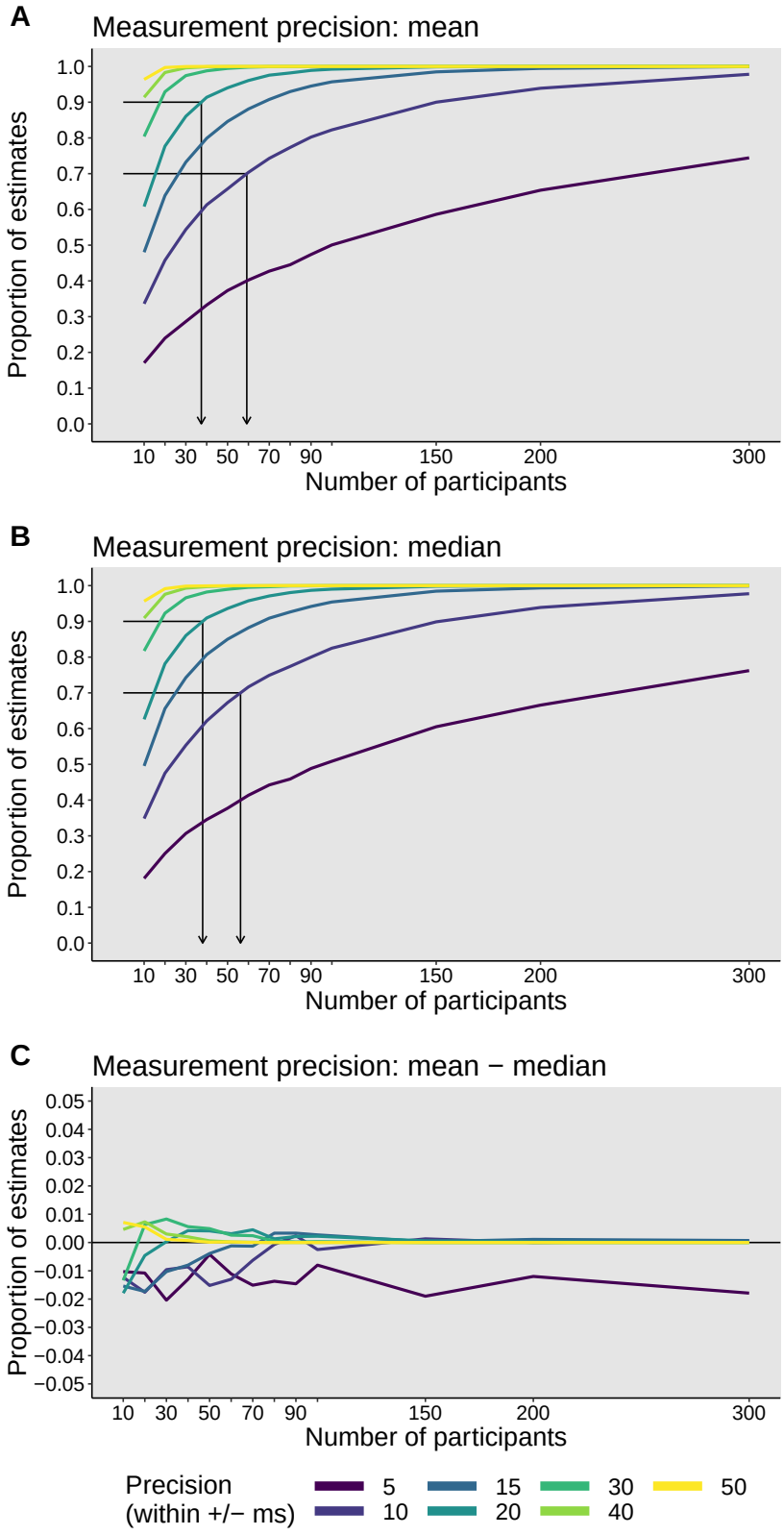

Figure 12. FLP dataset: group measurement precision for the difference between the Non-Word and Word conditions. Measurement precision was estimated by using a simulation with 10,000 iterations, 200 trials per condition and participant, and varying numbers of participants. Results are illustrated for the mean (**A**), the median (**B**), and the difference between the mean and the median (**C**).

conclusion: the recommendation by Miller (1988) to not use the median when comparing distributions that differ in sample size was ill-advised, for several reasons. First, the bias of the difference between medians is not due to differences in sample sizes but to differences in skewness, so that in practice, the bias can be small or even negligible. Second, the bias can be strongly attenuated by using a percentile bootstrap bias correction. However, although the bootstrap bias correction appears to work well in the long run, for a single experiment there is no guarantee it will provide an estimation closer to the truth. One possibility is to report results with and without bias correction. Third, the sample distributions of the mean and the median are positively skewed when sampling from skewed distributions, such that computing the mean of the sample distributions to estimate bias can be misleading. If instead we consider the median of the sample distributions, we get a better indication of the expected bias in a typical experiment, and this median bias tends to be smaller for the sample median than the sample mean. Fourth, the mean and the median differ very little in the measurement precision they afford. Thus, there seems to be no rational for preferring the mean over the median as a measure of central tendency for skewed distributions. If the goal is to accurately estimate the central tendency of a RT distribution, while protecting against the influence of outliers, the median is far more efficient than the mean (Wilcox & Rousselet, 2018). Providing sample sizes are moderately large, bias is actually not a problem, and the typical bias is actually very small. So, if we have to choose between the mean and the median, the median appears to be a better option.

Other aspects need to be considered. In particular, in an extensive series of simulations, Ratcliff (1993) demonstrated that when performing standard group ANOVAs, the median can lack power compared to other estimators.Ratcliff's simulations involved ANOVAs on group means, in which for each participant and each condition, very few trials (7 to 12) were available. These small samples were then summarised using several estimators, including the median. Based on the simulations, Ratcliff recommended data transformations or computing the mean after applying specific cut-offs to maximise power.

However, these recommendations should be considered with caution because the results could be very different with more realistic sample sizes. Also, standard ANOVA on group means are not robust, and alternative techniques should be considered, involving trimmed means, medians and M-estimators (Field & Wilcox, 2017; Wilcox & Rousselet, 2018). More generally, standard procedures using the mean lack power, offer poor control over false positives, and lead to inaccurate confidence intervals (Field & Wilcox, 2017; Wilcox & Rousselet, 2018).

Data transformations are not ideal either, because they change the shape of the distributions, which contains important information about the nature of the effects. Transformations can also fail to remove the skewness of the original distributions, and they do not effectively deal with outliers (Wilcox, 2017). Also, once data are transformed, inferences are made on the transformed data, not on the original ones, an important caveat that tends to be swept under the carpet when results are discussed.

Finally, truncating distributions introduces bias too, especially when used in conjunction with the mean (Miller, 1991; Ulrich & Miller, 1994).

Indeed, common outlier exclusion techniques lead to bias estimation of the mean (Miller, 1991). When applied to skewed distributions, removing any values more than 2 or 3 standard deviation from the mean affects slow responses more than fast ones. As a consequence, the sample mean tends to underestimate the population mean. And this bias increases with sample size because the outlier detection technique does not work for small sample sizes, which results from the lack of robustness of the mean and the standard deviation (Wilcox & Keselman, 2003). The bias also increases with skewness. Therefore, when comparing distributions that differ in sample size, or skewness, or both, differences can be masked or created, resulting in inaccurate quantification of effect sizes. Truncation using absolute thresholds (for instance by removing all RT $<$ 300 ms and all RT $>$ 1,200 ms and averaging the remaining values) also leads to potentially severe bias of the mean, median, standard deviation and skewness of RT distributions

(Ulrich & Miller, 1994). The median is much less affected by truncation bias than the mean though. Also, the median is very resistant to the effect of outliers and can be used on its own without relying on dubious truncation methods. In fact, the median is a special type of trimmed mean, in which only one or two observations are used and the rest discarded (equivalent to 50% trimming on each side of the distribution). There are advantages in using 10% or 20% trimming in certain situations, and this can done in conjunction with the application of t-tests and ANOVAs for which the standard error terms are adjusted (Wilcox & Keselman, 2003; Wilcox, 2017).

Overall, there isn't much convincing evidence against using the median of RT distributions, if the goal is to use only one measure of location to summarise the entire distribution. Clearly, a better alternative is to not throw away all that information, by studying how entire distributions differ (G. A. Rousselet et al., 2017; Heathcote et al., 1991). Whatever the approach chosen, we need to consider skewness at both levels of analysis: in each participant and across participants. Depending on the amount of skewness and spread at each level and for each condition, an important question is how to spend our money: by investing in more trials or in more participants (Rouder & Haaf, 2018)? An answer can be obtained by running simulations, either data-driven using available large datasets or assuming generative distributions (for instance exGaussian distributions for RT data). Simulations that take skewness into account are important to estimate bias and power. Assuming normality can have disastrous consequences (Wilcox & Rousselet, 2018). In many situations, simulations will reveal that much larger samples than commonly used are required to improve the precision of our measurements (Button et al., 2013).

Data and code availability

All the figures in this article are licensed CC-BY 4.0 and can be reproduced using notebooks in the R programming language (R Core Team, 2018) and datasets available on *figshare* (G. Rousselet & Wilcox, 2018). The main R packages used to generate the data and make the figures are *ggplot2* (Wickham, 2016), *cowplot* (Wilke, 2017), *tibble* (Müller & Wickham, 2018), *tidyr* (Wickham & Henry, 2018), *retimes* (Massidda, 2013), *knitr* (Xie, 2018), *HDInterval* (Meredith & Kruschke, 2016), and the essential *beepr* (Bååth, 2018).

References

Bååth, R. (2018). beepr: Easily play notification sounds on any platform [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=beepr (R package version 1.3)

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The annals of statistics*, 267–277.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., ... Pallier, C. (2010). The french lexicon project: Lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496.

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour research and therapy*, *98*, 19–38.

Heathcote, A., Popiel, S. J., & Mewhort, D. (1991). Analysis of response time distributions: An example using the stroop task. *Psychological Bulletin*, *109*(2), 340.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Massidda, D. (2013). retimes: Reaction time analysis [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=retimes (R package version 0.1-2)

Matzke, D., Dolan, C. V., Logan, G. D., Brown, S. D., & Wagenmakers, E.-J. (2013). Bayesian parametric estimation of stop-signal reaction time distributions. *Journal of Experimental Psychology: General*, *142*(4), 1047.

Meredith, M., & Kruschke, J. (2016). Hdinterval: Highest (posterior) density intervals [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=HDInterval (R package version 0.1.3)

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 539.

Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*, *43*(4), 907–912.

Müller, K., & Wickham, H. (2018). tibble: Simple data frames [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tibble (R package version 1.4.2)

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58.

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between stroop and simon effects using delta plots. *Attention, Perception, & Psychophysics*, *72*(7), 2013–2025.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological bulletin*, *114*(3), 510.

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223.

Rousselet, G., & Wilcox, R. (2018). Reaction times and other skewed distributions: problems with the mean and the median. *figshare*. doi: 10.6084/m9.figshare.6911924

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*(2), 1738–1748.

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, *123*(1), 34.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 475–482.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H., & Henry, L. (2018). tidyr: Easily tidy data with 'spread()' and 'gather()' functions [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tidyr (R package version 0.8.0)

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). San Diego, CA: Academic press.

Wilcox, R. R., & Keselman, H. (2003). Modern robust data analysis methods: measures of central tendency. *Psychological methods*, *8*(3), 254.

Wilcox, R. R., & Rousselet, G. A. (2018). A guide to robust statistical methods in neuroscience. *Current protocols in neuroscience*, *82*(1), 8–42.

Wilke, C. O. (2017). cowplot: Streamlined plot theme and plot annotations for 'ggplot2' [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=cowplot (R package version 0.9.2)

Xie, Y. (2018). knitr: A general-purpose package for dynamic report generation in r [Computer software manual]. Retrieved from https://yihui.name/knitr/ (R package version 1.20)