Behavior Research Methods
https://doi.org/10.3758/s13428-023-02084-7

Combining reaction-time distributions to conserve shape

Saul Sternberg[1]

Accepted: 3 February 2023

Abstract
To improve the estimate of the shape of a reaction-time distribution, it is sometimes desirable to combine several samples, drawn from different sessions or different subjects. How should these samples be combined? This paper provides an evaluation of four combination methods, two that are currently in use (the *bin-means histogram*, often called "Vincentizing", and *quantile averaging*) and two that are new (*linear-transform pooling* and *shape averaging*). The evaluation makes use of a modern method for describing the shape of a distribution, based on L-moments, rather than the traditional method, based on central moments. Also provided is an introduction to shape descriptors based on L-moments, whose advantages over central moments—less biased and less sensitive to outliers—are demonstrated. Whether traditional or modern shape descriptions are employed, the combination methods currently in use, especially bin-means histograms, based on averaged bin means, prove to be substantially inferior to the new methods. Averaged bin-means themselves are less deficient when estimating differences between distribution shapes, as in delta plots, but are nonetheless inferior to linear-transform pooling.

Keywords Vincentize · Quantile average · Reaction-time · Distribution · Linear-transform pooling · L-moments · Delta-plot

CONTENTS

The computations on which this report is based were conducted in R (R Core Team, 2022). Packages used include lmom (Hosking, 2019), lmomco (Asquith, 2021), robustbase (Maechler et al., 2022), SuppDists (Wheeler, 2022), and gamlss.dist (Stasinopoulos & Rigby, 2022).

✉ Saul Sternberg
saul@psych.upenn.edu

[1] University of Pennsylvania, Philadelphia, PA, USA

Published online: 30 May 2023

1. Introduction

Since the revival of interest in reaction-time (RT) measurement of the 1960s, the distributions of reaction-times, as well as their means, have been of interest. The shapes of observed distributions have been used to guide the choice of theoretical distributions (e.g., Christie & Luce, 1956; Hohle, 1965; McGill, 1963; Snodgrass et al., 1967; Sternberg, 1964; Stone, 1960). And inferential methods have been developed that depend on the properties of observed distributions, differences between them, and combinations of them, with no commitment to a particular theoretical distribution.

Examples include skewness and kurtosis (Vickers, 1979), the survivor interaction contrast (Townsend & Nozawa, 1995; Schweickert et al., 2012; Little et al., 2017), the summation test (Roberts & Sternberg, 1993; Schweickert et al., 2012), the race model inequality (Colonius & Vorberg, 1994; Gondan & Vorberg, 2021; Lombardi et al., 2019; Miller, 1982), the delta plot (de Jong et al., 1994; Ellinghaus & Miller, 2018; Mackenzie et al., 2022; Miller & Schwarz, 2021; Schwarz & Miller, 2012), tests of RT mixtures (Reynolds & Miller, 2009; Yantis et al., 1991), and the "short RT" and "long RT" properties (Sternberg, 1973; Vorberg, 1981; Townsend & Ashby, 1983, Ch. 8), and others (Schweickert & Giorgini, 1999). And it has been argued that the shapes of RT distributions are important for model discrimination (Ratcliff & Smith, 2010, p. 90). Some of the insights that have depended on distributional analysis, insights that were not evident from mean RTs alone, are summarized by Balota and Yap (2011). Descriptors of distributional shape and the quality of their estimators have therefore become important.

The *shape* of a distribution consists of all its features other than its location and its scale (or spread). Thus, for constants $a > 0$ and $b > 0$, the distributions of X and of $Y = a + bX$ differ in scale and location, but have the same shape: X and Y are then members of the same *location-scale family* (LSF).

Two major features of the shape of a distribution are its *skewness* (a measure of asymmetry: how much more the right side of the distribution is spread out than the left side) and its *kurtosis* (a measure of "tailweight": how much of the distribution is located in its tails versus its central region[1]). These features can be represented by a point in a plane in which the x- and y-axes represent measures of skewness and kurtosis, respectively. Given equal locations and scales, this is not to say that the shapes of two distributions with the same skewness and kurtosis are identical; their shapes may have other features that differ. However, a necessary condition for a particular theoretical distribution to be compatible with an observed distribution is that their skewness and kurtosis agree.

To estimate the shape of a reaction-time distribution when there is no convincing basis for knowing its form, it is often desirable to combine two or more samples of RTs (for example, RTs from different subjects in the same experimental condition), because the sizes of the individual samples are too small to provide accurate information about the shapes of their parent distributions. Such samples may come from parent distributions that differ only in location and/or scale, as might occur for several reasons, which include the effects of practice, differences between subjects, differences between stimuli, whether the previous trial's stimulus is the same or different, or vocal response differences such as those due to differences among words in delays associated with the movements of the speech articulators and in the time taken for a voice detector to be triggered.[2] Combining such samples by simple pooling is inappropriate: Doing so is likely to result in shape distortion.

Suppose, for example, that sixty RTs are collected from the same condition in each of two sessions, and we wish to combine the RTs from the two sessions for a subject. Suppose further that the mean RTs differ, a practice effect. If we knew that the distributions of which the two sets of RT are samples were members of the same LSF, then we might choose to combine the samples by quantile averaging (discussed by Thomas & Ross, 1980, and by Jiang et al., 2004) with the goal of preserving their shape, or by pooling after adjusting their means and scales.

Unfortunately, we do not know that these distributions are members of the same LSF. If they were, the Q-Q plot of the samples would be linear, but the sample sizes are too small

[1] The meaning of "kurtosis", when it is defined, traditionally, as the standardized fourth central moment, is unclear (Westfall, 2014). The interpretation is clearer when it is defined in terms of L-moments, discussed below and in Appendix A. The corresponding density functions in Panels A and B of Fig. D1 have the same values of L-skewness, but differ in L-kurtosis, with values of L-kurtosis smaller in Panel B. The light curve in Panel A has about the same L-kurtosis as the broken curve in Panel C (0.175 vs. 0.172), but differs in L-skewness (.100 vs .172).

[2] Such speech detection variations can be as great as 100 ms, larger than the word- frequency effect often of primary interest (See Table 2 in Andrews & Heathcote, 2001, and Table 4 in Spieler & Balota, 1997). Detection delays can be reduced by using a speech detector that applies appropriately different but low thresholds to high and low audio frequencies, along with filtering that rejects the brief non-speech lip noises that exceed the low thresholds. Alternatively, estimates of such effects can be used to correct the RTs before combining data from different words. Without either suitable instrumentation or such correction, the shapes of distributions of combined RTs may be misleading.

to confidently determine the linearity of such a Q-Q plot.[3] Also, it is argued in Section 5.3 that it is hard to justify the assumption that practice (as an example of one factor whose levels might differ between samples to be combined) influences only the location and scale, leaving the shape invariant.

What should we require of a good combination method? A minimum requirement is that if the parent distributions of the samples to be combined differ in location and scale but have the same shape, that shape should be approximated by the combination. A second requirement is that if the parent distributions differ in shape (which seems more likely), the combination method should produce a distribution whose shape features (among which are, e.g., skewness and kurtosis) are, in some sense, averages of the shape features of the components. (This is why the title of the present paper refers to "conserving" rather than "preserving" shape.)

In this paper, I report on the evaluation of four combination methods, two of which are in common use. Because of their bias and variability for small samples and their sensitivity to outliers, traditional measures of shape, based on central moments, can be problematic when used for this purpose. As demonstrated in Appendices B and C, these difficulties are mitigated when the measures of shape are based on L-moments, introduced by Hosking (1990).[4]

2. Traditional measures of shape: standardized central moments

For a distribution of x, the traditional measures of skewness and kurtosis are functions of its central moments ("C-moments", sometimes called "product moments"), $\mu_r = E\left(x-\bar{x}\right)^r$, ($r$=2, 3,...), with skewness measured by $\widehat{\beta}_1 = m_3/m_2^{3/2}$ and kurtosis measured by $\widehat{\beta}_2 = m_4/m_2^2$, where m_k is an estimate of μ_k.[5,6] Thus the skewness and kurtosis of an observed distribution can be represented as a point in the β_1–β_2 plane, and these properties of the shapes permitted by any particular family of distributions may be represented in that plane as a point for some families, or a curve, or a region, for others.[7] For several examples, see Johnson et al. (1994), and Pearson (1963). By comparing the location of the $\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ point for an observed distribution with the point, curve, or region permitted by a given family, one can learn whether the observations can be described by a member of that family, or ways in which members of that family would be inadequate.

Generation of skewness–kurtosis curves To show how such an implicit curve is generated from the parametric equations that specify β_1 and β_2, consider the ex-Gaussian distribution, familiar to some psychologists, as an example. For that distribution, whose density function is specified by Eq. (5), the parametric equations are $\beta_1 = 2\tau^3/(\sigma^2+\tau^2)^{1.5}$ and $\beta_2 = 3+6\tau^4/(\sigma^2+\tau^2)^2$. A little algebra, aided by defining $r = \tau^2/(\sigma^2+\tau^2)$, leads to the curve: $\beta_2 = 3+6(\beta_1/2)^{4/3}$.

Among the difficulties associated with using β_1 and β_2 is that not only are their estimates biased, especially in small samples (Appendix B; Headrick, 2011, pp. 2, 13; Marzuki et al., 2012; Vogel & Fennessey, 1993) and to an extent that depends on the underlying distribution, but, also, because they depend on C-moments (for which the RTs must be exponentiated), the variability of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ is of special concern, and (as exemplified in Appendix C) they are especially sensitive to outliers.[8]

3. Modern measures of shape: standardized L-moments

These difficulties are mitigated by the use of L-moments, $\{\lambda_r\}$ (instead of C-moments) and their standardized forms, sometimes called "L-moment ratios", especially L-skewness ($\lambda_{s3} = \lambda_3/\lambda_2$) and L-kurtosis ($\lambda_{s4} = \lambda_4/\lambda_2$).[9] Introduced by

[3] Thomas and Ross (1980) and Jiang et al. (2004) showed that quantile averaging has desirable properties when the parent distributions of the components belong to the same LSF, and Thomas and Ross (1980) recommended the use of Q-Q plots to test this. However, I have seen no evidence in the psychological literature of this test having been applied as a prelude to quantile averaging. Furthermore, the evaluations described below show that with realistic sample sizes, quantile averaging is inadequate even when the samples are from members of the same LSF.

[4] For an example of applying L-moments in the analysis of the shapes of RT distributions, see Sternberg & Backus, 2015.

[5] Because of the standardization by m_2, β_1 and β_2 are dimensionless quantities: The same (β_1, β_2) point represents distributions that can differ in location and/or scale.

[6] Traditionally, β_1 is defined as μ_3^2/μ_2^3, and the skewness measure as $\sqrt{\beta_1}$ (Kendall & Stuart, 1969, Chapter 3). For the purpose of this paper, β_1 will be defined as $\mu_3/\mu_2^{3/2}$.

[7] I use the term "family of distributions", or just "family" to indicate a theoretical distribution with the values of its parameters unspecified, and the term "distribution" to denote a particular member of that family, with specified parameter values.

[8] The variability of these measures, emphasized for C-moments of RT data by Ratcliff (1979) and Heathcote et al. (1991), increases with the variance of the underlying distribution. Thus, when their use is anticipated, experimental methods are called for that reduce this variance, such as extensive practice, rest periods, adequate warning signals, and performance incentives and feedback.

[9] L-skewness and L-kurtosis have often been represented by τ_1 and τ_2, respectively, by analogy to β_1 and β_2. To avoid their being confused with τ, often used for the exponential parameter of the ex-Gaussian distribution, popular in psychology, this paper will use λ_{s3} and λ_{s4}, etc., instead, where the "s" indicates the standardization.

Hosking (1990), L-moments have been the subject of substantial theory and numerous applications, especially in hydrology, where a primary use is to help select and fit distributions that will aid in predicting the risk of extreme weather events (see, e.g., Asquith, 2011; Headrick, 2011; Hosking, 1992, 2006; Hosking & Wallis, 1997; Royston, 1992; Selaman et al., 2007; Ulrych et al., 2000; Vogel & Fennessey, 1993; Wang, 1996). The first six L-moments are defined and discussed in Appendix A. Estimates of L-moments and their ratios are "nearly unbiased for all underlying distributions" (Vogel & Fennessey, 1993, p. 1746; Headrick, 2011, Section 3), and, because L-moments are linear functions of sample values (hence the "L"), they are less sensitive to outliers than C-moments (An et al., 2023).[10] Comparisons of the bias and the sensitivity to outliers of $(\widehat{\lambda}_{s3}, \widehat{\lambda}_{s4})$ and $(\widehat{\beta}_1, \widehat{\beta}_2)$ for several members of the Wald (inverse-Gaussian) and ex-Gaussian families are provided in Appendix B and C.[11] The bias is substantial for $(\widehat{\beta}_1, \widehat{\beta}_2)$, especially for small samples and for members of the families for which skewness and kurtosis are high, but negligible for $(\widehat{\lambda}_{s3}, \widehat{\lambda}_{s4})$.[12] In all cases, the effects of an outlier are considerably smaller, proportionally, for $\widehat{\lambda}_{s3}$ and $\widehat{\lambda}_{s4}$ than they are for $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

Just as for the traditional measures β_1 and β_2, one can represent features of the shape of an observed distribution as a point in the λ_{s3}—λ_{s4} plane, and consider its location relative to the points, curves, or regions in that plane associated with various distribution families.[13] Because of the desirable properties of L-moments and their ratios, this will be the approach taken in the present paper in considering the shapes of RT distributions. The primary application of L-moments has been to select theoretical distributions appropriate for sets of data, for which λ_{s3} and λ_{s4} have been employed. However, to describe the shape of a distribution, supplementing these values with λ_{s5} and λ_{s6} is useful.

The points and curves in Fig. 1 and similar figures that follow come from several sources, including the R packages *lmom* (Hosking, 2019) and *lmomco* (Asquith, 2021) and the appendix of Hosking and Wallis (1997), which provides L-moments for several distributions as well as approximations to some of the curves that relate λ_{s4} to λ_{s3} (also available as Table 10.1 in Asquith, 2011). For others I used distribution simulations with large samples and wide ranges of parameter values, determined the resulting estimates of $\widehat{\lambda}_{sk}$, $k = 3, 4, 5, 6$, which provided the equivalent of the required parametric equations, and plotted $\widehat{\lambda}_{s4}$ against $\widehat{\lambda}_{s3}$, and $\widehat{\lambda}_{s6}$ against $\widehat{\lambda}_{s5}$, as the resulting implicit curves.

3.1. Why four L-moment ratios?

Many applications of L-moment ratios (standardized L-moments) make use only of L-skewness (λ_{s3}) and L-kurtosis (λ_{s4}), because these are usually sufficient to decide on which of several theoretical distributions might best fit a set of observations.[14] However, λ_{s5} and λ_{s6} also differ among alternative theoretical distributions (see, e.g., Fig. 1B), and can be important for at least two reasons. First, interesting different distributions associated with the same point in the λ_{s3}—λ_{s4} plane exist (see Karvanen & Nuutinen, 2008). Second, they permit additional tests of goodness of fit of a theoretical distribution to a sample. And third, Headrick (2011) has adapted the "power method" of approximating distributions with specified central moments to doing so for distributions with specified L-moments, λ_1 and λ_2 and L-moment ratios, λ_{sk}, $(k = 3, 4, 5, 6)$. His procedure, which makes use of polynomials in powers of logistic or Gaussian random variables with coefficients that are functions of these quantities, can be used to generate random variables with the specified L-moments and L-moment ratios, and to specify the corresponding density functions.[15] For an adequate specification of distributional shape, Headrick's power method requires λ_{s5} and λ_{s6}, as well as the lower-order L-moments. It is easy to show that the resulting distribution is altered in response to changes in λ_{s5} and λ_{s6}, indicating that all six quantities are necessary.

4. Four combination methods

The methods to be described can be applied to any number of samples. They are *bin-means histograms, quantile averaging, linear-transform pooling* and *shape averaging*.[16] Both

[10] When considering the sums of stochastically independent random variables, it should be noted that (functions of) L-moments do not share the very useful additivity property of the cumulants of a distribution (κ_r), which are functions of the C-moments. For example, when x and y are stochastically independent, then, for their sum, $x + y$, the cumulants are additive: $\kappa_r\,(x + y) = \kappa_r\,(x) + \kappa_r\,(y)$, $(r \geq 2)$. It follows, for example, that the variance, third moment, and higher cumulants of the duration of a sequence of stochastically independent iterated processes all increase linearly with the number of iterations.

[11] These two distributions are among those that have been applied to reaction-time data; see, e.g., Luce (1986), Anders et al. (2016).

[12] These levels of bias for L-skewness and L-kurtosis can be further reduced by adopting the estimates developed by Withers & Nadarajah (2011).

[13] Examples of distributions that are represented by regions, rather than points or curves, are the four-parameter Kappa distribution (Hosking & Wallis, 1997; Kjeldsen et al., 2017) and the generalized lambda distribution (Asquith, 2007).

[14] But see, e.g., Asquith, 2014; Hosking, 2007; Karvanen & Nuutinen, 2008.

[15] To derive the density function of the distribution of combined samples when using the shape averaging combination method, or to generate random samples from the combined distribution, Headrick's procedure (or an equivalent) would be necessary.

[16] I have not included the method devised by Cousineau et al. (2016), because it requires estimating the minimum RT, which depends on assuming a feature of the shape of the distribution.

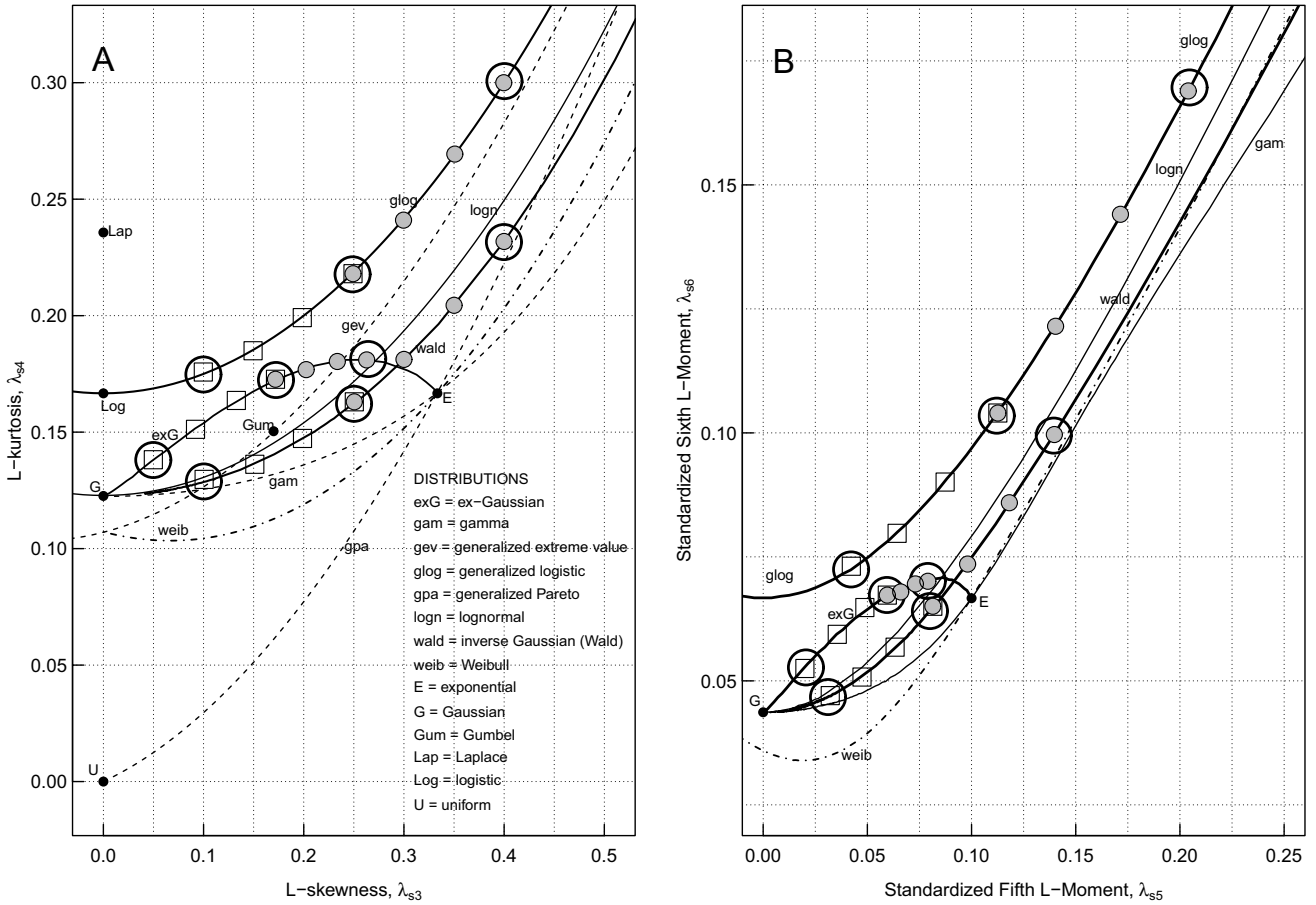

Fig. 1 Distribution Families and Test Distributions. Panel **A** The λ_{s3}-λ_{s4} (L-skewness - L-kurtosis) plane with points and curves shown for fourteen distribution families. Values of L-skewness and L-kurtosis for the three test distributions in each of the generalized logistic, ex-Gaussian, and Wald families are indicated by large open circles located on their respective curves. The open squares (for Mix.1) and filled circles (for Mix.2) mark points on each of these curves for the six sets of four distributions used in tests of combining samples from distributions with different shapes (See Tables 1, 2, and 3). Panel **B** The λ_{s5}-λ_{s6} plane with curves for six distribution families, including the generalized logistic, ex-Gaussian, and Wald families. Open circles, open squares and filled circles have the same meaning as in Panel **A**

of the first two, which are currently in use to combine RT samples, have unfortunately been called by the same name: "Vincentizing" (Ratcliff, 1979; Rouder & Speckman, 2004; Thomas & Ross, 1980).[17]

4.1. Bin-means histograms

This method was devised by Ratcliff (1979) who named it "Vincentizing"; its use was advocated by Dawson (1988) and Heathcote (1996), who independently provided programs to implement it, in the context of fitting the ex-Gaussian distribution. Select a desired number, m, of bins. To ensure that the sample, containing n observations, can be partitioned into m equal-size bins, replicate the sample m times. (The probability distribution of a replicated sample is the same as the probability distribution of the sample.) Define a set of equally spaced proportions, p_1, p_2,..., p_{m+1}, where $p_1 = 0$ and $p_{m+1} = 1$. Let $q_j = q(p_j)$ be the quantile of the replicated sample associated with proportion p_j. This defines a set of m bins, $\{b_k\}$, $(k = 1, 2,, \ldots, m)$, with boundaries (q_k, q_{k+1}), each containing values $\{X_{ki}\}$, $(i = 1, 2,.., n)$. For each sample, determine the means, $\{X_{k\bullet}\}$ of the n values in each bin.[18,19]

[17] Referring to these two methods, Jiang et al. (2004, p. 187), who distinguish them, claimed that "In practice, the two variants yield highly similar results." In what follows we shall see that one can be substantially inferior to the other, especially if the number of bins in bin-means histograms is as small as has been recommended. See footnote 21.

[18] The R function I used to obtain bin means is provided in Appendix G.

[19] The midpoint of the j^{th} bin is $(p_j + p_{j+1})/2$. It has sometimes been claimed that the j^{th} bin mean is the quantile associated with the j^{th} bin midpoint. This is (approximately) true if and only if the distribution of observations within the j^{th} bin is (approximately) symmetric. Thus, bin means for different samples are quantiles whose associated probabilities are likely to differ.

The estimate of the combined distribution of several samples is achieved in two steps: First the number of bins, m, is selected, each sample is partitioned into m bins as described, the bin means $\{X_{k\bullet}\}$ are obtained for each sample, and corresponding bin means are averaged over samples, yielding the averaged bin means, $\overline{X_{k\bullet}}$, $(k = 1, 2,, \ldots, m)$. Second, the averaged bin means are used to create a *bin-means histogram*, representing the combined distribution, that consists of m equal-area rectangles, extending from $\overline{X_{k\bullet}}$ to $\overline{X_{k+1\bullet}}$, $(k = 1, 2,, \ldots, m)$. The number of bins typically used has ranged from five to 20. In the present paper, the method is evaluated for two variants, with $m = 15$ and $m = 30$.[20, 21]

In some applications, the combined distribution is represented by the averaged bin means themselves, usually obtained from five or ten bins, by plotting them as a function of bin number, for example, or by plotting differences between corresponding averaged bin means associated with different experimental conditions as a function of bin number. Such "delta plots" are discussed in Section 7.

4.2. Quantile averaging

A set of equally spaced proportions from 0.0 to 1.0 is defined. In my evaluation of this method, they were separated by 0.01, 0.02, or 0.05, but only my findings for 0.01, which provided the best estimates, are reported. For each sample, the quantiles associated with the set of proportions are determined. The samples are combined by averaging corresponding quantiles.[22,23]

In quantile averaging, each quantile is associated with a specified member of a set of equally spaced proportions; the sets of quantiles that are averaged across samples are associated with the same set of proportions. In bin-means averaging, a bin mean is the mean of the observations between two such quantiles; the proportion for which it is the quantile is between the two corresponding proportions, but exactly where depends on the distribution of observations in that bin. Thus, in contrast with quantile averaging, the sets of bin means that are averaged across samples are likely to be associated with different sets of proportions.

4.3. Linear-transform pooling

Suppose we have observations $\{RT_{1j}\}$ and $\{RT_{2j}\}$ from two distributions, F_1 and F_2 that have the same shape, but differ in location and scale. Let F_1 and F_2 have locations loc_1 and loc_2 and scales $scale_1$ and $scale_2$. If we eliminated the differences between these locations and scales by linear transformations and pooled the distributions, then the distribution of the combined samples F_{12} would have the same shape as the components. We could do this by defining a target location, $locT$, and a target scale, $scaleT$, and adjusting the distributions as follows, for $k = 1, 2$:

$$RT_{kj.adj} = locT + \left(RT_{kj} - loc_k\right)\left[\frac{scaleT}{scale_k}\right]. \tag{1}$$

If we applied these transformations to samples from the two distributions, and pooled the transformed values, the combined sample would have F_{12} as its parent population, with the same shape as F_1 and F_2, as desired. One problem with this idea is that we almost never know the locations or scales of the parent distributions; we have only their estimates, $\widehat{loc}_1$, $\widehat{loc}_2$, $\widehat{scale}_1$, and $\widehat{scale}_2$, obtained from the samples. Replacing the unknown values by their estimates, the transformations are, for $k = 1, 2$:

$$RT_{kj.adj} = locT + \left(RT_{kj} - \widehat{loc}_k\right)\left[\frac{scaleT}{\widehat{scale}_k}\right] \tag{2}$$

The shape of the combined sample is not affected by the choices of the targets $locT$ and $scaleT$; in the evaluations of this method, I have used the means of the corresponding sample estimates.

Now, suppose that the samples come from distributions with different shapes. In that case, the shape of the pooled adjusted values should differ from the shapes of both of the components, but should be a kind of average of these shapes, an average whose properties will be determined in evaluations of the method.

By combining each of five alternative estimates of location with each of six alternative estimates of scale I generated 30 variants of this method. The location measures are the mean, the median, the 20% winsorized mean, the 20%

[20] The bin means have sometimes been called "Vincentiles" (Van Zandt, 2000). They are also quantiles, of course, as are any values within the bounds of the sample, but these quantiles are not defined in terms of the proportions to which they correspond.

[21] According to Ratcliff (1979, p. 459), whose examples included nine bins (Fig. 2) and 24 bins (Fig. 4), the procedure can be used with "as few as ten observations per subject". According to Heathcote (1996), who recommends between 4 and 19 bins, and whose example (Fig. 7) has 19 bins (p. 435), "For good ex-Gaussian fits, Vincentizing requires as few as 20 observations per condition". He also suggests (p. 436) that "Vincent averaging combines data across subjects or conditions without distorting the shape of the underlying distribution." According to Andrews & Heathcote (2001, p. 517), "Average vincentiles approximately preserve the shape of underlying distributions for each participant, as long as the individual participants' distributions are unimodal and smooth." In what follows, the validity of these claims is tested.

[22] I used Hyndman & Fan's (1996) type-8 quantile estimator. This method might be improved by using a better quantile estimator, such as the one proposed by Navruz & Ozdemir (2020).

[23] This method (which they call "Vincentizing") has been thoroughly discussed by Jiang et al. (2004), Rouder & Speckman (2004), and by Thomas & Ross (1980), who also discuss variants that make use of transformed observations. The differences between quantile histograms and bin-means histograms (sometimes called "Vincent histograms") are discussed by Van Zandt (2000). See also footnote 2 in Andrews & Heathcote (2001).

trimmed mean, and the mean of the 0.25 and 0.75 quantiles.[24] The scale measures are the second L-moment (λ_2), Qn (a robust measure; Rousseeuw & Croux, 1993),[25] the interquartile range, the median absolute deviation (*mad*), the standard deviation, and the square root of the 20% winsorized variance (Wilcox, 2017).[26] As described in Appendix E, comparison of the estimation precision of the thirty variants led to the choice of the mean and Qn as the best measures of location and scale, respectively.

4.4. Shape averaging

Perhaps the simplest way to specify the shape of the combined samples is to determine the four corresponding shape measures, λ_{skj}, $(k=3, 4, 5, 6)$, for each of the samples, and to average them. For two samples the results would be $\lambda_{sk\bullet}=(\widehat{\lambda}_{sk1}+\widehat{\lambda}_{sk2})/2$, $(k=3, 4, 5, 6)$. For deciding whether the combined shape is approximately consistent with any particular theoretical distribution, this is sufficient. If samples from a distribution with the resulting specified shape, or its density function, are desired, they can be generated using polynomial transformations of Gaussian or logistic random numbers (Headrick, 2011), where the polynomials are determined by $\widehat{\lambda}_1$, $\widehat{\lambda}_2$, and the $\widehat{\lambda}_{sk}$, $(k=3, 4, 5, 6)$.

Most of the distributions shown in Fig. 1 are represented by curves rather than points, which means that parameter changes are associated with shape changes. Consider a distribution family that is represented as a curve on such a plot. Members of that family with different parameters will, in general, appear as separate points on its curve. The point that represents their mean L-skewness and mean L-kurtosis will not fall on the curve, so cannot represent the shape of a member of the same family (the curves in the λ_{s3}–λ_{s4} and λ_{s5}–λ_{s6} planes that represent the best-known distributions are all nonlinear, and, except for the ex-Gaussian, concave up). However, parts of many of the curves are close to linear, so that the shape of a combined distribution whose shape measures are close to the means of the component shape measures, will be similar, but not equal, to a member of the same family. For example, consider samples from two or more members of the ex-Gaussian family with different σ/τ ratios, corresponding to different points on the ex-Gaussian curve. The point that represents their average shape will lie below the curve because it is concave down, but not very far from the curve, because the concavity is not great.

[24] This last measure was used by Wolfe et al. (2010), who called it the "x- score transform".

[25] Qn is given by the .25 quantile of the distances $\{|x_i - x_j|; i<j\}$

[26] The linear-transform method was used by Roberts and Sternberg in their "summation test" (1993; Sternberg & Backus, 2015, Appendix), but using the median and the interquartile range as measures of location and scale, respectively, which are not the choices that minimize the mean estimation error in my simulations (Appendix E).

4.5. Bypassing sample combination by assuming a theoretical distribution

An alternative approach to combining samples that is sometimes taken (and is advocated by Rouder & Speckman, 2004) is to assume a particular theoretical form for the RT distribution (often the ex-Gaussian, by psychologists), and to fit that distribution to each sample. Having done so, one can make inferences based on the fitted parameter values averaged over two or more such samples. Rouder and Speckman (2004) show that when the distribution from which samples are drawn is known, such parameter averaging is sometimes (but not always) preferable to quantile averaging followed by fitting the known theoretical distribution to the quantile averages. However, without a strong argument that specifies the theoretical distribution, this makes sense only after first showing that the shapes of the sample distributions are consistent with that distribution, which often requires that the samples be combined to permit valid shape measures to be obtained, using a method that doesn't depend on a commitment to a particular distribution, such as one of those under consideration.

4.6. Some details

Information loss in the four methods As illustrated in Appendix F, information about the distribution of observations less than the smallest bin mean and greater than the largest is not represented in the bin-means histogram. Also, creating the distribution representations for both of the first two methods entails the loss of shape-relevant information: the locations of the values within each bin (in bin-means histograms) and their locations between successive quantiles (in quantile averaging). More information is lost with fewer (larger) bins, and with greater separation between successive quantiles. The linear-transform method entails no such information loss. To the extent that different distributions can have the same values of λ_{sk}, $(k=3, 4, 5, 6)$, the shape averaging method loses shape-relevant information.

Estimation of L-moments of combined distributions L-moments are provided directly by the shape averaging method. Because L-moment estimators are based on means of linear functions of order statistics of small samples of various sizes, such estimates of the combined distribution require representing it in the form of observations, rather than bin means or quantiles. Such observations are provided directly by linear-transform pooling, but an additional step is required for quantile and bin-means averaging. Samples from the combined distributions created by these methods were computed by means of a deterministic version of inverse transform sampling. For quantile averaging, I created a fixed number of samples, equally spaced between each pair

Table 1 Generalized logistic distributions: parameter values and shape measures

Name	Mix.1	Mix.2	μ	α	κ	λ_{s3}	λ_{s4}	λ_{s5}	λ_{s6}
*glo.10	g1		396.0	24.0	−0.10	0.100	0.175	0.043	0.073
glo.15	g1		394.0	23.0	−0.15	0.149	0.185	0.064	0.080
glo.20	g1		392.0	22.0	−0.20	0.199	0.199	0.087	0.090
*glo.25	g1	g2	390.0	19.0	−0.25	0.250	0.219	0.113	0.104
glo.30		g2	388.0	21.0	−0.30	0.300	0.241	0.140	0.121
glo.35		g2	387.0	19.0	−0.35	0.351	0.269	0.171	0.144
*glo.40		g2	390.0	12.0	−0.40	0.400	0.300	0.204	0.169

Table 2 Shifted wald distributions: parameter values and shape measures

Name	Mix.1	Mix.2	μ	λ	θ	λ_{s3}	λ_{s4}	λ_{s5}	λ_{s6}
*Wald.10	w1		216.9	5042.0	183.1	0.100	0.129	0.032	0.047
Wald.15	w1		142.2	142.1	258.0	0.151	0.136	0.047	0.051
Wald.20	w1		105.8	584.0	294.0	0.199	0.147	0.063	0.057
*Wald.25	w1	w2	82.0	272.2	318.0	0.249	0.162	0.080	0.064
Wald.30		w2	66.3	144.0	334.0	0.300	0.181	0.098	0.074
Wald.35		w2	54.7	81.0	345.0	0.350	0.204	0.118	0.086
*Wald.40		w2	45.9	47.7	354.1	0.400	0.232	0.140	0.101

of quantiles. For quantiles separated by probability differences of 0.01 the number of samples was 5. Similarly, for bin-means averaging, I created a fixed number of samples, equally spaced between each pair of successive bin means. For analyses that make use of 15 and 30 bins, the numbers of samples were 20 and 10, respectively. Appendix G provides the R code for the function used for this purpose.

5. Evaluation of combination methods

I conducted three kinds of evaluation of the four methods and their variants.

Ten same-shape parent distributions. In one kind I combined ten samples from distributions of the same shape, but with differences in location and scale.
Four same-shape parent distributions. In a second kind, I combined four samples from distributions of the same shape, but with differences in location and scale.
Four different-shape parent distributions. In a third kind, I combined four samples from distributions in the same theoretical family, but with differences in shape, as well as differences in location and scale.

In each case I used the same four sample sizes: 30, 60, 100, and 200. It can be argued that the third kind of evaluation corresponds more closely to what we would expect with real data, where shape is unlikely to be invariant (see Section 5.3). On the other hand, a good method should work well when the parent distributions of the samples to be combined have the same shape, and where the measure of success of the shape estimate is therefore more straightforward than when parent shapes differ.

5.1. Test distributions

For the first two kinds of evaluation, I chose nine basic distributions to provide samples. The names of these distributions are marked by asterisks in Tables 1, 2, and 3. Six were selected to span a large range of skewness and kurtosis values: three generalized logistic distributions, whose density function is:

$$f(t;\kappa,\mu,\alpha) = \left[\frac{\exp\{-(1-\kappa)Y\}}{\alpha(1+\exp\{-Y\})^2}\right], \tag{3}$$

where $Y = -(1/\kappa)\ \log\{1-\kappa(t-\mu)/\alpha\}$, if $\kappa \neq 0$, and $Y = (t-\mu)/\alpha$, if $\kappa = 0$, and three shifted Wald (inverse Gaussian) distributions (Anders et al., 2016; Miller et al., 2017) with shift parameter θ, whose density function is:

$$f(t;\lambda,\mu,\theta) = \left(\frac{\lambda}{2\pi T^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda(T-\mu)^2}{2\mu^2 T}\right\}, \tag{4}$$

where $T = t - \theta$.

Because of its popularity among psychologists, I also included three ex-Gaussian distributions, whose density function is

$$f(t;\mu,\sigma,\tau) = (1/\tau)\exp\left\{-\frac{(t-\mu)}{\tau}+\frac{\sigma^2}{2\tau^2}\right\}\Phi\left\{\frac{(t-\mu)}{\sigma}-\frac{\sigma}{\tau}\right\}, \quad (5)$$

where Φ is the distribution function of the standard Gaussian distribution. The shapes of each of these distribution families can be represented by a curve in the four-dimensional shape space with dimensions $\{\lambda_{sk}\}$, $k = 3, 4, 5, 6$; any member of the family is represented by a point on this curve. All distributions were adjusted to have standard deviations of approximately 45 units, and means of approximately 400 units, corresponding to millisecond units for typical RTs, before their locations and scales were modified in the evaluations. The distributions used for the third kind of evaluation are indicated by "Mix.1" and "Mix.2" in Tables 1, 2, and 3. For the mixture indicated by "g1" in Table 1, for example, one sample was drawn from each of the first four generalized logistic distributions in the table, for each replication. As shown in the three tables, there were six such sets of four distributions, g1, g2, w1, w2, e1, and e2. The three tables show parameter values and shape measures, in order of increasing L-skewness (λ_{s3}). Note that the names of the distributions include approximations of λ_{s3} (the value of L-skewness).

For the basic Wald and generalized logistic distributions I chose parameters to provide L-skewness values of approximately 0.10, 0.25, and 0.40. This was not possible for the ex-Gaussian distribution, because of its more limited range of shapes. Density functions of the nine basic distributions (those marked by asterisks in Tables 1, 2, and 3) are displayed in Appendix D.

Panels A and B of Fig. 1 are projections of the shape space onto the first and second pair of dimensions, respectively. The nine basic distributions are represented by large open circles; the six sets of four distributions (g1, g2, w1, w2, e1, and e2) are represented by smaller open squares and filled circles, marking points on the curves for the generalized logistic, ex-Gaussian, and Wald distributions.

5.2. Evaluation procedure

It is helpful to think of the shape of the combined distribution produced by a particular method as an estimate of the shape of a parent distribution. The best method is one for which such estimates are the most accurate.

Samples from same-shape parent distributions In one set of tests, for each of the nine parent distributions I adjusted the location and scale by a linear transformation to create ten distributions of the same shape with ten different locations (adding 0, 10, . . . , 90 units to the parent location) and ten different scales (with multipliers from 0.55 to 1.45 applied to the scale of the parent), and with the location and scale adjustments paired randomly in each of the 500 replications. In each replication, then, the ten distributions were members of the same LSF. In each replication, a sample of size N, was obtained from each of the resulting distributions, providing ten samples of equal size, where $N = 30$, 60, 100, and 200. Each of the combination methods was then applied to each of the resulting sets of ten samples.

A second set of tests was similar to the first, except for the number of samples combined: For each of the nine parent distributions I created four different distributions of the same shape with four different locations (by adding 0, 30, 60, and 90 units to the location) and four different scales (by applying four scale multipliers from 0.55 to 1.45 to the scale of the parent), again with the location and scale adjustments paired randomly in each of 500 replications. In each replication a sample of size N was obtained from each of the resulting distributions, providing four samples of equal size, where $N = 30, 60, 100$, and 200. Each of the combination methods was then applied to each of the resulting sets of four samples.

Samples from different-shape parent distributions In a third set of tests, the samples were drawn from parent distributions with different shapes, such as the four different shapes called "g1" in Table 1: $\{\Lambda_{skn}\}$, $(k = 3, 4, 5, 6; n = 1, 2, 3, 4)$. For present purposes, the mean shape of the distributions is the shape that corresponds to their four mean L-moments, in this case, $\{\Lambda_{sk\bullet}\}$, $(k = 3, 4, 5, 6)$. The mean Euclidean distance and bias of the combination shape relative to the mean shape are the measures I chose for evaluation.

5.3. Implausibility of shape invariance

Before proceeding further, let us consider the likelihood that the samples to be combined from actual experiments are members of the same LSF. In a typical case, these samples might be generated by the same subject in different sessions, and influenced by practice, or might be from different subjects, influenced by individual differences. If the differences are limited to the *locations* of the parent distributions of these samples, then they will have the same shape. But suppose that in addition to the location of the distribution, its *variance* (μ_2) is influenced by practice, or differs among individuals.

Recall that, in terms of the central moments, skewness and kurtosis are defined, respectively, as $\beta_1 = \mu_3/\mu_2^{3/2}$, and $\beta_2 = \mu_4/\mu_2^2$. Suppose that practice causes μ_2 to be reduced so that $\mu_2^* = \alpha\mu_2$, where μ_2 and μ_2^* are, respectively, the

Table 3 Ex-Gaussian distributions: parameter values and shape measures

Name	Mix.1	Mix.2	μ	σ	τ	λ_{s3}	λ_{s4}	λ_{s5}	λ_{s6}
*exg.05	e1		374.0	37.1	25.5	0.050	0.138	0.020	0.053
exg.09	e1		368.0	32.1	31.5	0.092	0.151	0.036	0.059
exg.13	e1		364.0	27.6	35.5	0.133	0.164	0.049	0.065
*exg.17	e1	e2	362.0	23.4	38.4	0.172	0.172	0.059	0.067
exg.20		e2	360.0	20.0	40.3	0.202	0.177	0.066	0.068
exg.23		e2	358.0	16.5	41.9	0.234	0.180	0.073	0.070
*exg.26		e2	357.0	13.1	43.0	0.264	0.181	0.079	0.070

second central moments before and after practice, and $0 < \alpha < 1$. For the distributions before and after practice to have the same shape, both β_1 and β_2 must be invariant. Given the change in μ_2, this requires that practice also causes quantitatively appropriate changes in in μ_3 and μ_4: In particular, whatever the value of α, we must also have $\mu_3^* = \alpha^{3/2}\mu_3$ and $\mu_4^* = \alpha^2\mu_4$. To me, such corresponding differences among the central moments seem implausible unless there is either persuasive evidence for shape invariance, or some confirmed theoretical reason to expect it.[27] We should therefore require a combination method to work well when samples are drawn from parent distributions that differ in shape. However, although the combining of samples from same-shape distributions may be unrealistic, we would also expect any good combination method to be effective in this simplified situation, which has the advantage that the measure of estimation accuracy of the combined distribution is straightforward: its shape should approximate the common shape of the parent distributions.

5.4. Measures of success for same-shape parent distributions

What would it mean for a combination method to be successful? In the case of actual samples, the corresponding parent distributions, whose shapes we care about, are unknown. However, for the simulations that were used to test these methods, the shapes of the parent distributions are known and specified. If the parents have the same shape, we can ask how well the shape of the combination approximates that shape.

For replication j, let X_{1j}, X_{2j},.., X_{nj} represent the n samples, and let X_{cj} represent their combination, arrived at by one of the methods under consideration. In each test I determined, for each replication, four standardized L-moments of X_{cj}, λ_{cskj}, ($k = 3, 4, 5, 6$), and compared them to the corresponding shape measures of the parent distribution, Λ_{sk}, ($k = 3, 4, 5, 6$). The *mean estimation error* is the Euclidean distance in the four-dimensional shape space between these two points, determined for each replication, and averaged over the 500 replications. The *estimation bias* is the Euclidean distance between the *mean* estimate of the shape of the combination, $\lambda_{csk\bullet}$ and the corresponding shape measures of the parent distribution, Λ_{sk}, ($k = 3, 4, 5, 6$). The goals were to find the combination method for which the mean estimation error as well as the estimation bias are minimized.

5.5. Measures of success for different-shape parent distributions

As shown in Tables 1, 2, and 3, tests of combining samples from different-shape parent distributions all involved four different parent distributions from the same theoretical family, with shapes $\{\Lambda_{skm}\}$, ($k = 1, 2, 3, 4$, $m = 1, 2, 3, 4$). For each replication, samples of equal size, n, were drawn from each of these parents, with $n = 30, 60, 100, 200$. As above, for replication j, let X_{1j}, X_{2j}, X_{3j}, and X_{4j} represent the four samples, and let X_{cj} represent their combination. Again, in each test and for each replication, j, I determined four L-moment ratios of X_{cj}, λ_{cskj}, ($k = 3, 4, 5, 6$). In this case I compared the shape of X_{cj} to the *mean shape* of the four parent distributions, $\Lambda_{sk\bullet}$, ($k = 3, 4, 5, 6$). The mean estimation error and the estimation bias were then determined as in the case of same-shape parent distributions.

6. Evaluation results

6.1. Mean estimation error

For the reasons described in Appendix E, the variant of the linear-transform pooling method that I decided to use for comparison with the other methods is the one with the mean as the location measure, and Qn as the scale measure. And,

[27] A contrasting view is expressed by Ratcliff and Smith (2010, p. 90), according to whom "Invariance of distribution shape is one of the most powerful constraints on models of RT distributions.That the diffusion model predicts this invariance is a strong argument in support of its use in performing process decomposition of RT data." See Sternberg & Backus (2015) for a discussion of arguments for (and evidence against) shape invariance.

as mentioned above, for bin-means histograms, the results for both fifteen and thirty bins are reported. Mean estimation error is shown in Fig. 2, for combining ten samples (Panel A) and four samples (Panel B) from same-shape distributions, and four samples from different-shape distributions (Panel C). Quantile averaging and bin-means histograms are consistently worse than the other methods. When the samples being combined are from different-shape distributions, the advantages of the linear and mean-shape methods over quantile averaging are increased for $N = 100$ and $N = 200$. Unsurprisingly, the estimation error is reduced by increases in sample size.

6.2. Magnitude of the estimation bias

The magnitude of the estimation bias for a combination method is the Euclidean distance between the *mean* estimated shape produced by that method, averaged over the 500 replications, and either the (common) shape of the parent distributions (when they have the same shape), or their *mean* shape (when they differ in shape). The magnitude of the estimation bias is shown in Fig. 3, for combining ten samples (Panel A) and four samples (Panel B) from same-shape distributions, averaged over the nine parent distributions, and four samples from different-shape distributions (Panel C), averaged over the six sets of four different-shape distributions. The effect of sample size on bias magnitude is substantially smaller for the mean shape and linear combination methods.

6.3. Nature of the estimation bias

We have considered the magnitude of the estimation bias in Fig. 3; it is also important to consider its form. This is shown for samples of size $N = 60$ in Fig. 4 (for combining ten samples from same-shape parent distributions), in Fig. 5 (for combining four samples from same-shape parent distributions), and in Fig. 6 (for combining four samples from different-shape parent distributions). I chose to report the results for this sample size because it seems to be the most likely of the four sample sizes to be used for combining samples. Although the magnitude of the bias is sensitive to sample size, as we have seen, its direction tends not to be.

Estimates using the mean quantile method and based on combining samples with same-shape parents are more strongly biased for $\widehat{\lambda}_{s4}$ and especially $\widehat{\lambda}_{s6}$ than for $\widehat{\lambda}_{s3}$ and $\widehat{\lambda}_{s5}$. Unlike bin-means histograms, when such estimates are based on data from different-shape parents, L-skewness ($\widehat{\lambda}_{s3}$) is over-estimated.

Estimates of all four shape measures based on bin-means histograms are substantially smaller than their target values, enough so that they are consistent with the wrong family of distributions. For example, estimates based on data from the generalized logistic distribution would lead one to conclude that the underlying family of distributions is the lognormal or ex-Gaussian, estimates based on data from the ex-Gaussian distribution would suggest the gamma or Weibull family, and estimates based on data from the Wald distribution would suggest the Weibull distribution. These findings suggest that some conclusions based on past applications of bin-means histograms may have to be reconsidered.

All four methods have difficulty estimating the shape of the distribution with the greatest L-skewness and L-kurtosis; see especially the estimates of the variant of the generalized logistic distribution with the greatest values of λ_{s3} and λ_{s4} in Fig. 4 and 5. However, the problem is not due to high skewness alone, as the biases of estimates of the most skewed Wald distribution, based on all except bin-means histograms, are not great.

6.4. Evaluation of combination methods using traditional shape measures

As mentioned in Section 1, and documented in Appendices B and C, traditional shape measures, based on central moments, are inferior to L-moments in some ways. Perhaps for this reason, they were not used for evaluating the effectiveness of bin-means histograms by its inventors and promotors (Ratcliff, 1979; Heathcote et al., 1991), who argued against the use of central moments. Although inferior to L-moments (which were not available in 1979), central moments are hardly worthless for such purposes, and, indeed, have some useful properties not shared by L-moments.[28] It therefore seemed of interest to compare the effectiveness of the alternative methods for which this was possible, using traditional shape measures (the shape averaging method had to be excluded because its results are expressed in terms of L-moments). What would have been the outcome of evaluations of these methods before the invention of L-moments? This was done for combinations of ten samples from same-shape parent distributions. For each replication, j, and each of the methods other than shape averaging, I estimated β_{1j} (traditional skewness) and β_{2j} (traditional kurtosis) of the combined distribution, X_{cj}, using unbiased estimators of the central moments, as follows: Let $y_j = X_{cj} - \overline{X_{cj}}$. Then $m_{2j} = \left[n/(n-1)\right]\overline{y_j^2}$, $m_{3j} = \left[n^2/(n-1)(n-2)\right]\overline{y_j^3}$, and $m_{4j} = \left[n^2(n+1)/(n-1)(n-2)(n-3)\right]\overline{y_j^4}$. Finally, $\widehat{\beta}_{1j} = m_{3j}/m_{2j}^{3/2}$, and $\widehat{\beta}_{2j} = m_{4j}/m_{2j}^2$. I then determined the mean estimation error (the mean over replications of the Euclidean distance in the β_1–β_2 plane between these values, and the corresponding values for the parent distributions).

[28] See Footnote 10, and, for a recent validation of their use for inferring distributional shape, Section 6.3 in Sternberg (2016).

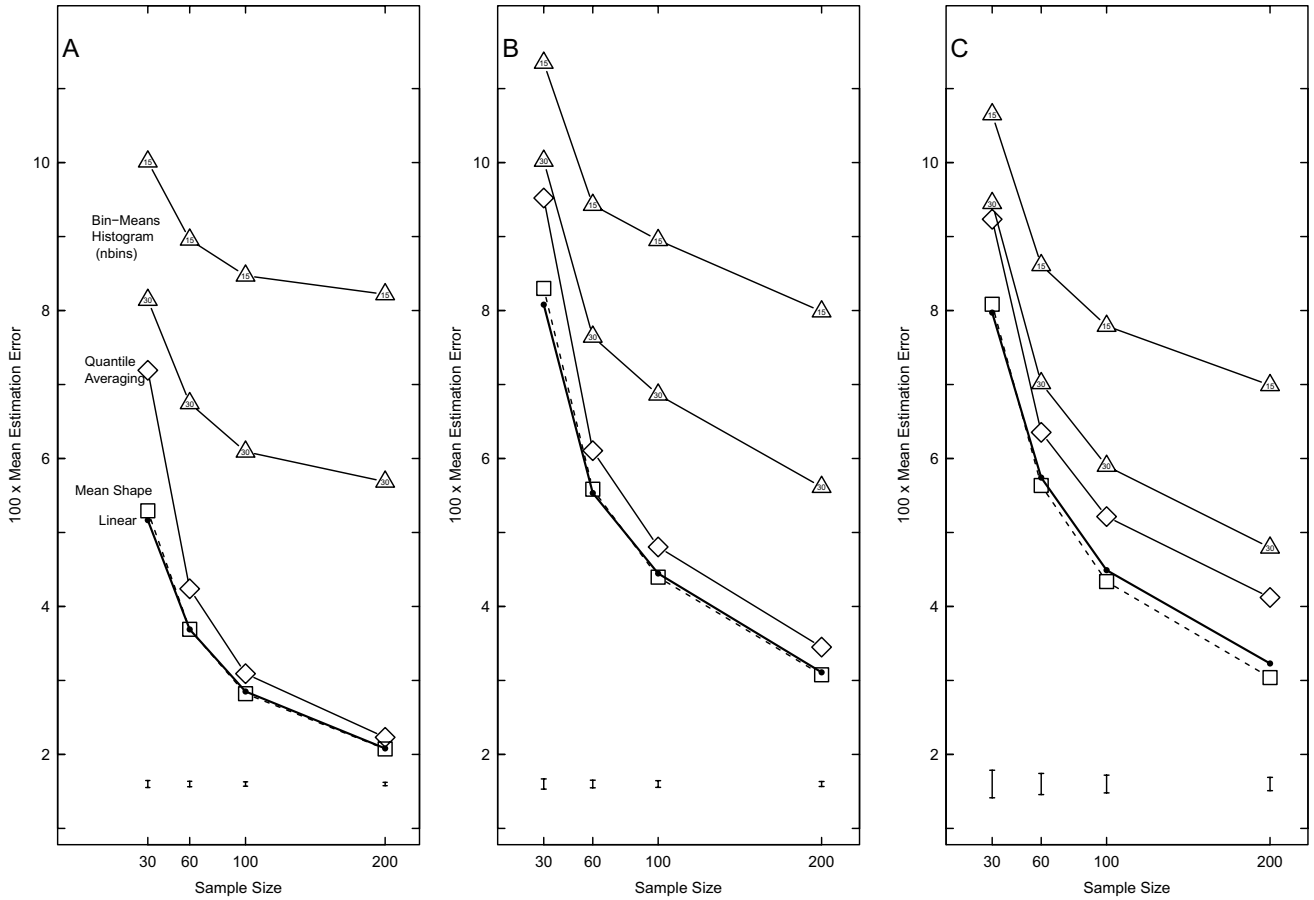

Fig. 2 Mean Estimation Error as a Function of Sample Size for Five Combination Methods. Panels **A** and **B** give mean estimation error as a function of sample size when the samples being combined are from same-shape parent distributions, ten samples in Panel **A**, four samples in Panel **B**. Panel **C** gives the results when four samples from different shape parent distributions are combined. Also shown for each sample size are the mean values of $\pm SE$, averaged over the five methods. The SEs vary relatively little across methods, from 93 to 113% of the mean for ten samples, and from 91 to 109% of the mean for four samples

Figure 7A shows the means of these estimation errors, averaged over the 500 replications as well as the nine parent distributions, and should be compared to Fig. 2A; Using shape estimates based on the third and fourth central moments, the advantage of the linear-transform method over quantile averaging is relatively greater than when using shape estimates based on four L-moments. I also determined the mean magnitude of the estimation bias—the Euclidean distance in the $\beta_1-\beta_2$ plane between the *mean* estimated shape and the corresponding values for the parent distribution, averaged over the nine distributions; these values are shown in Fig. 7B, which should be compared to Fig. 3A. Again, the $\beta_1-\beta_2$ measures reveal a relatively greater advantage of the linear-transform method over quantile averaging for the larger samples.

7. Estimation of shape differences

Relations between the shapes of RT distributions, rather than the shapes themselves, are sometimes of interest. For example, one might want to compare shapes of RT distributions under two conditions to determine whether differences are associated with just the longer RTs. One approach is to estimate the difference between the distribution functions of two distributions, possibly after using linear transformations to equate their locations and scales. This kind of analysis leads to the question of how close the estimated shape difference between two distributions is to the true shape difference, using alternative methods of shape estimation. If the biases were approximately equal for the distributions to be compared, then we might be fortunate, as they would "cancel out" in the estimated difference.

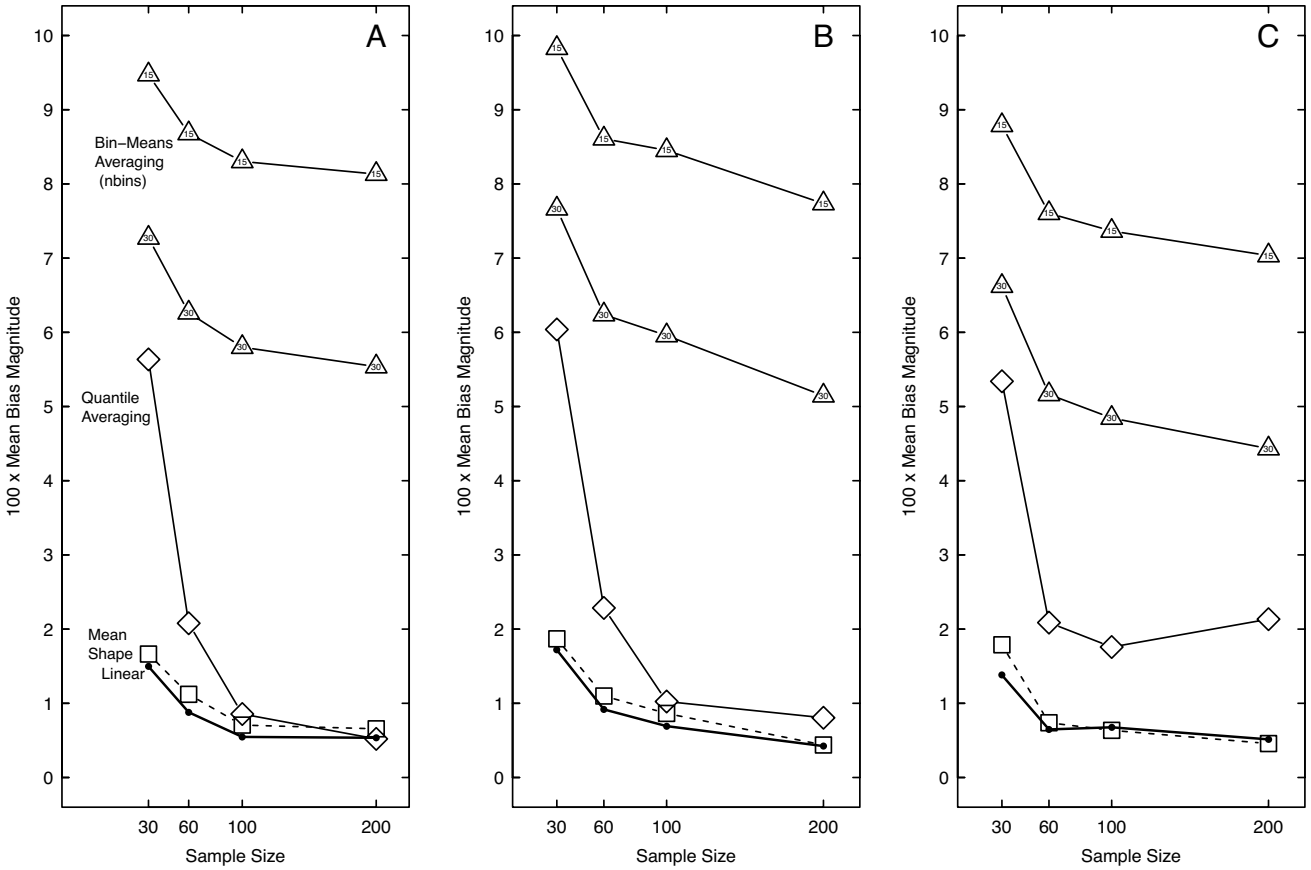

Fig. 3 Mean Magnitude of Estimation Bias as a Function of Sample Size for Five Combination Methods. Panel **A** gives the results when samples from ten distributions with the same shape are combined, Panel **B** when samples from four distributions with the same shape are combined, and Panel **C** when samples from four distributions with different shapes are combined

I used the nine basic distributions (those with asterisks in Tables 1, 2, and 3), determining the true shape difference for each of the 36 pairs of different distributions. These were compared to the differences among the corresponding pairs of estimated shapes of the combinations of ten samples of size $N = 60$ from same-shape parent distributions, using the five combination methods. In each case I computed the Euclidean distance between the estimated and true shape difference in the four-dimensional shape space defined by standardized L-moments. The results are shown in Table 4. The mean, maximum, and minimum Euclidean distance over distribution pairs, multiplied by 100, are shown for each method, first for all 36 distribution pairs, then for the 21 pairs that remain when the two most highly skewed distributions, glo40 and wald40 (which produce the greatest distances) are removed, and finally for the three pairs that involve just the three ex-Gaussian distributions.

Among the alternative methods, linear-transform pooling provided the best estimates of shape differences, and bin-means histograms with 15 bins provided the worst estimates.

An important alternative approach to comparing distributions is the computing of "delta plots" (Schwarz & Miller, 2012, pp. 556–558), first used by De Jong et al. (1994), which typically makes use of bin-means averaging with five or ten bins to combine distributions over subjects for each of two conditions. Delta plots use bin-means averages themselves, rather than the bin-means histograms derived from them, to compare two distributions. An example is provided by Balota et al. (2008, Experiment 1, Fig. 6), who assessed the effects of semantic priming on word-naming RTs by comparing bin-means averages of the RT distributions with related versus unrelated primes. Using ten bins, they measured the priming effect as the set of differences between bin-means averages under primed and unprimed conditions. The finding that these estimated differences were approximately equal for the ten bin-means averages was interpreted to mean that the shapes of the RT distributions under the two conditions are approximately the same. A second example is provided by Johnson et al. (2012), whose analysis of the effect of a transposed-letter neighbor on the RT for word naming included examining the

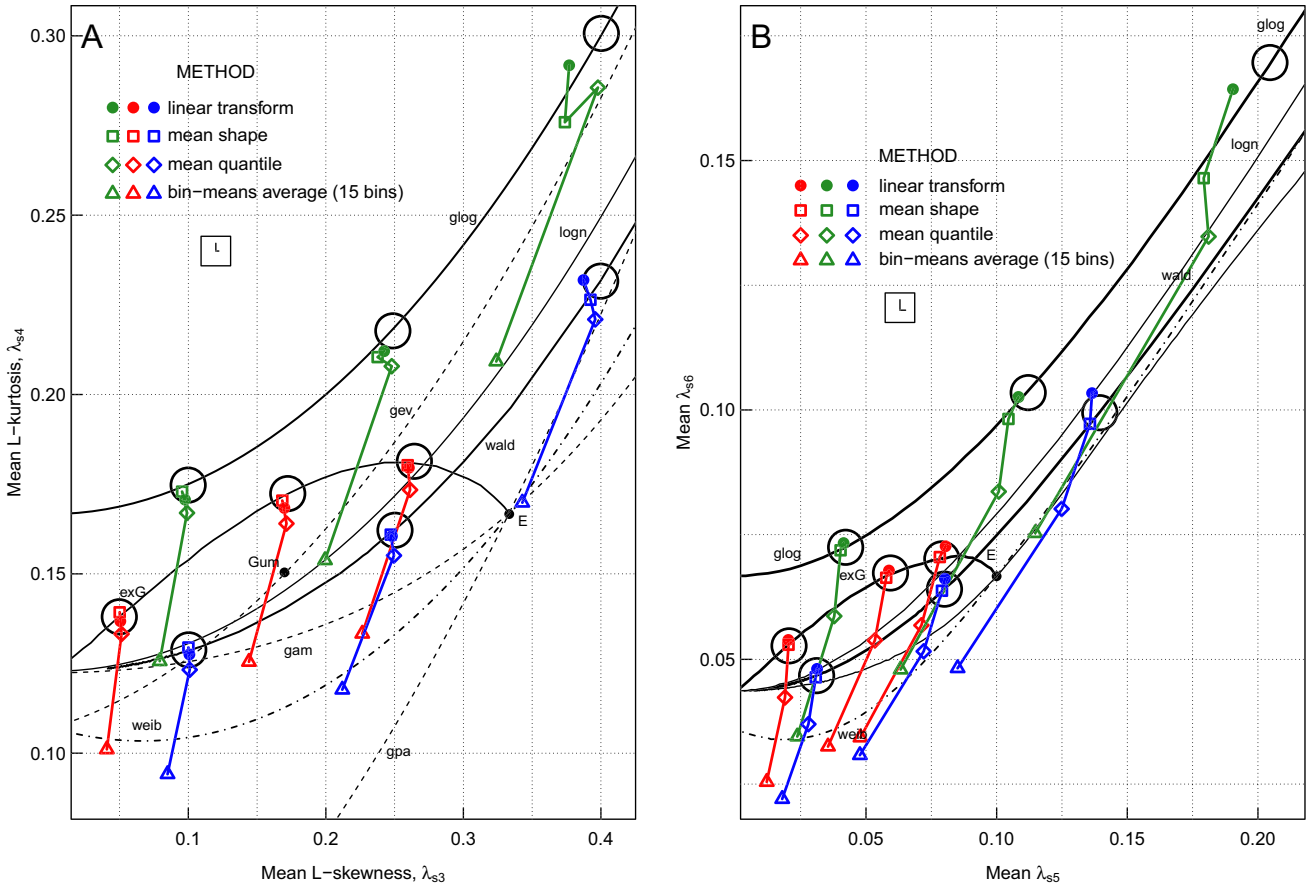

Fig. 4 Estimates provided by four combination methods in the $\lambda_{s3} - \lambda_{s4}$ plane (Panel **A**) and the $\lambda_{s5} - \lambda_{s6}$ plane (Panel **B**) when samples of size 60 from ten same-shape distributions are combined. The large circles represent the true shapes of the nine test distributions. Estimates of the same distribution are connected by line segments. Estimates of members of the three different distribution families are colored differently to aid discrimination. Lengths of limbs of the "L" under the legend represent the maximum SE of the estimates, along the x- and y-axes

set of differences between bin-means averages, also with ten bins, on trials with such a neighbor versus trials without one (their Fig. 3), and noting that the presence of a neighbor tended to influence the longer RTs selectively, producing a difference in shape.

Accuracy of delta plots It is of interest to ask about the accuracy of delta plots compared to other ways of comparing two distributions. Consider first the accuracy of bin-means averages as estimates of the quantiles of individual distributions For the nine basic distributions we have been considering (marked by asterisks in Tables 1, 2, and 3, all positively skewed) the estimation errors, averaged over 100 replications, are substantial only for the rightmost bin, which represents the right-hand tail of the distribution. For these distributions, estimation errors of the midpoint quantile (corresponding to probability = 0.90) by the fifth bin-mean average of five bins ranged from 6.4 to 19.2 ms, with a mean of 12.0 ms; and estimation errors of the midpoint quantile (corresponding to probability = 0.95) by the tenth bin-mean average of ten bins ranged from 5.1 to 24.3 ms, with a mean of 11.7 ms. Corresponding estimation errors provided by linear-transform pooling were smaller: For the 0.90 quantile, they ranged from – 0.9 to 4.7 ms with a mean of 0.7 ms, and for the 0.95 quantile they ranged from – 6.8 to 7.9 ms with a mean of – 0.2 ms.

To consider the accuracy of delta plots I selected three pairs of distributions, drawn from the nine basic distributions: exg.05 versus exg.26, exg.05 versus glo.40, and Wald.10 versus Wald.40. For each of the five distributions in these comparisons, I created ten samples of size 60 with different locations and scales, and used three procedures

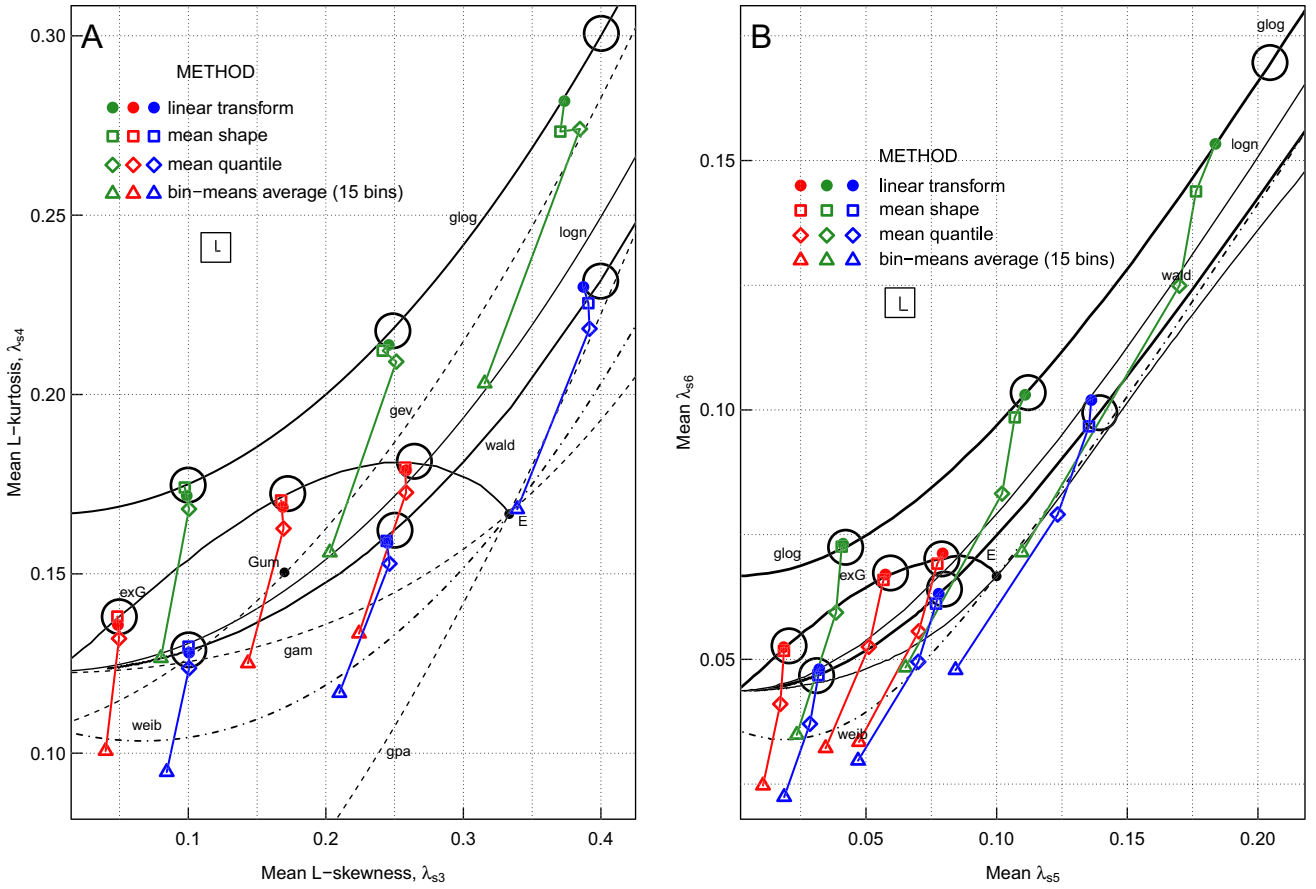

Fig. 5 The same as Fig. 4, except that samples from four same-shape distributions are combined

to combine the samples: linear-transform pooling, bin-means averaging with 10 bins, and bin-means averaging with five bins. Figure 8 shows the true differences together with estimates based on means from 100 replications. The differences between bin-means averages are plotted against the midpoints of the corresponding bins. Estimating differences between the right-hand tails of the distributions appears to cause difficulties for bin-means averaging: the differences are underestimated by a mean of 8.9 ms (for the fifth of five average bin means) and by 10.6 ms (for the tenth of ten average bin means). Corresponding mean bias values of the 0.90 and 0.95 quantiles provided by linear-transform pooling (midpoints of the corresponding bins) are overestimates with much smaller magnitudes: 0.5 and 1.1 ms, respectively.

Precision of delta plots In addition to the bias of an estimation method it is important to know about its precision. Consider first the precision of estimates of quantiles of individual distributions. To compare the precision of the continuous measure produced by linear-transform pooling to the precision of averaged bin means (five or ten discrete values associated with bin midpoints), I chose quantiles of the distributions produced by the former that corresponded to the bin midpoints. For example, for five bins, the bin midpoints are the probabilities 0.1, 0.3, 0.5, 0.7, and 0.9, so I chose the five corresponding quantiles. With 100 replications, we thus had 100 sets of each of five (ten) values for each method and each of the nine basic distributions. Averaging over distributions, the mean standard deviation for bin-means averaging with five (ten) bins was 2.53 (2.69) ms; for the corresponding quantiles estimated by linear-transform pooling, the corresponding values were 2.29 (2.41) ms. For the rightmost bin these values were 5.53 (8.41) ms and 4.25 (6.22) ms. Thus, estimates obtained by linear transform pooling are somewhat more precise than those obtained by bin-means averaging. These differences in precision for individual distributions are also shown by estimates of differences between distributions: Averaging standard deviations over the five (ten) difference values for each of the three comparisons shown in Fig. 8, the mean standard deviation of the bin-mean differences was 3.74 (4.01) ms; while the corresponding value for linear-transform pooling was 3.52 (3.77) ms, slightly smaller.

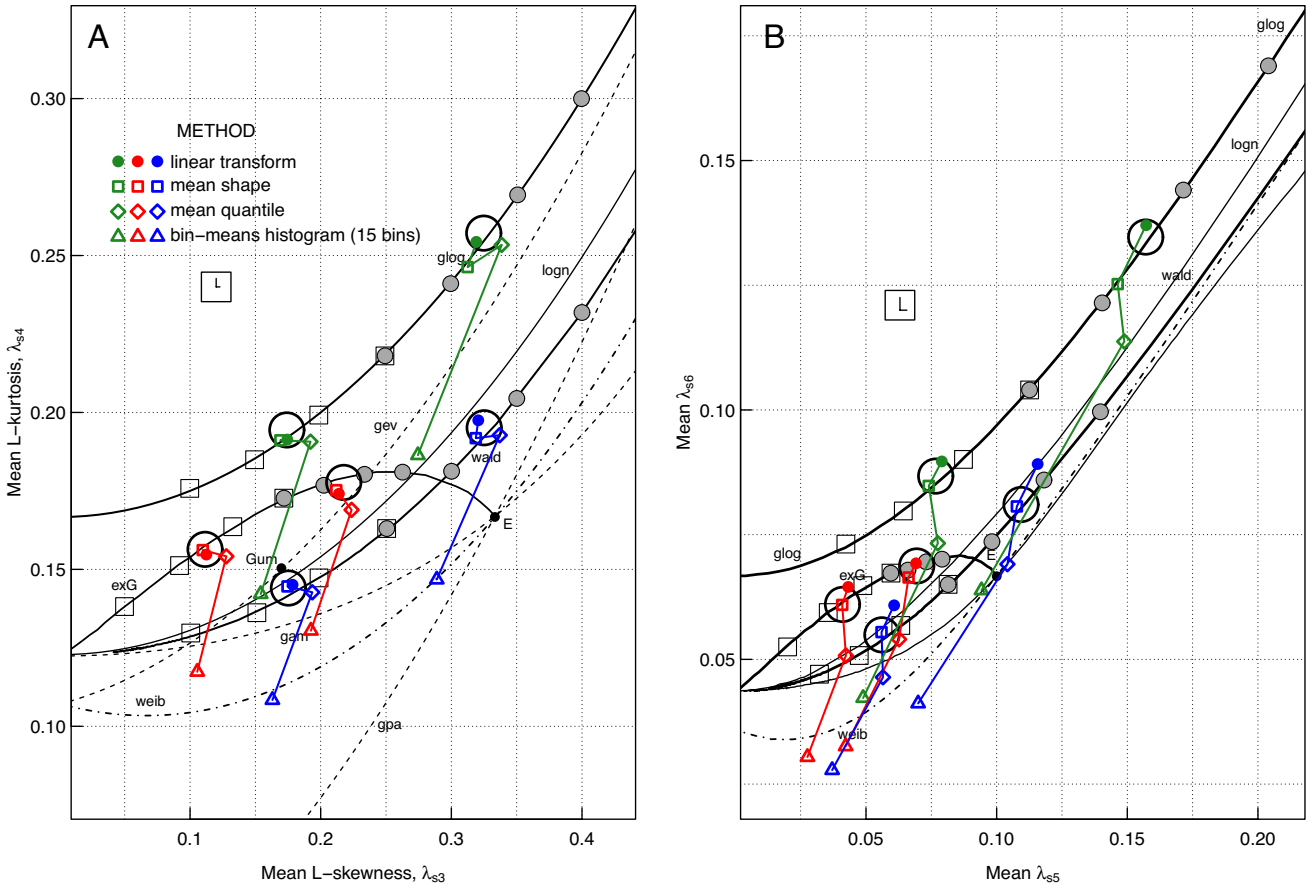

Fig. 6 Estimation Bias, Samples from Four Different − Shape Distributions Combined. The same as Fig. 4, except that samples from four different-shape distributions are combined, and each open circle represents the mean shape of one of the six sets of four distributions that provide those samples

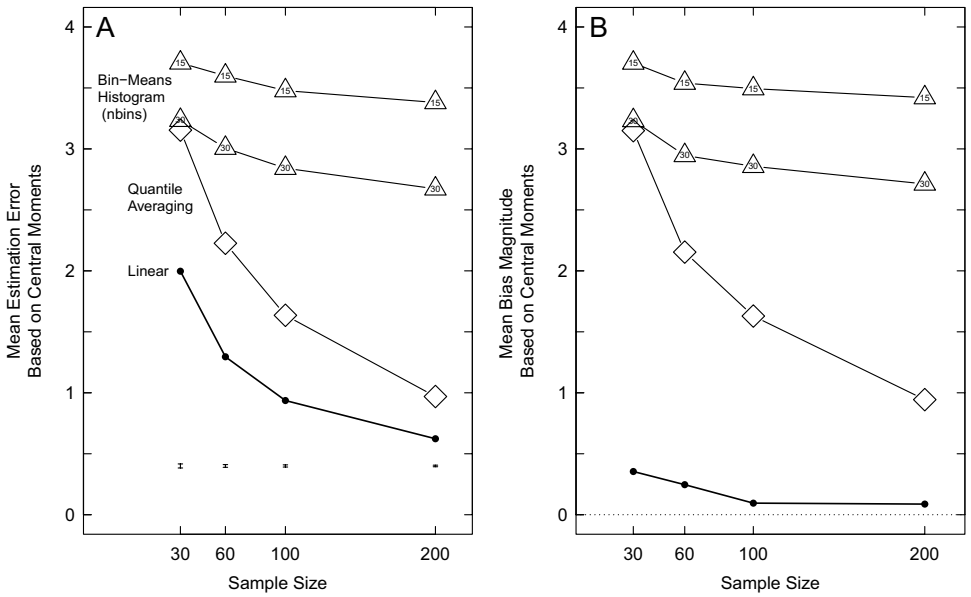

Fig. 7 Evaluation of Four Combination Methods in $\beta_1-\beta_2$ Shape Space. Mean estimation error (Panel **A**) and estimation bias magnitude (Panel **B**) as functions of sample size for four combination methods, with shape measures (β_1 and β_2) based on central moments. These results, for combining ten samples from same-shape parent distributions, show that neither the evaluation of these combination methods nor its qualitative results depend on the availability of L-moments as shape measures

Table 4 100×Euclidean distances between estimated and true shape differences

Method	36 Pairs			21 Pairs			3 Pairs		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Bin-Means Histogram (15 bins)	4.97	0.46	13.64	2.97	0.46	6.91	2.63	1.22	3.92
Bin-Means Histogram (30 bins)	3.68	0.31	10.31	2.17	0.31	5.19	1.88	0.83	2.79
Quantile Average	1.22	0.18	3.48	0.65	0.18	1.47	0.63	0.29	0.90
Mean Shape	1.51	0.65	5.07	0.63	0.07	1.69	0.44	0.20	0.56
Linear-Transform Pooling	1.07	0.42	2.96	0.45	0.04	0.93	0.45	0.39	0.52

Even with ten bins, detail is lost relative to what is provided by linear-transform pooling, whose accuracy and precision both appear to be better. Why restrict estimates to a small discrete set of points when linear-transform pooling provides a continuum, and with greater accuracy and precision? Based on this set of differences of three pairs of distributions, linear-transform pooling is to be preferred.

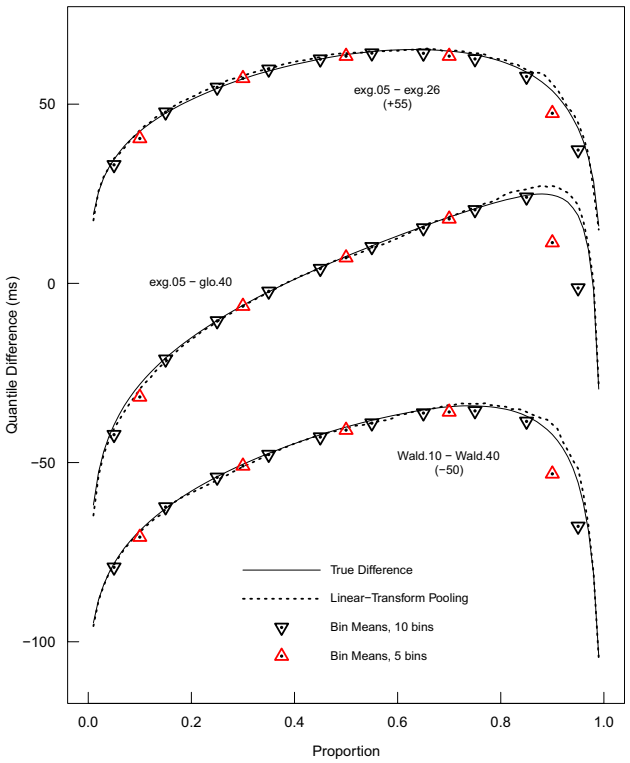

Fig. 8 Distribution – Difference Measures Compared. Values of the differences between members of three pairs of distribution functions and means of 100 replications of estimates of those values based on three combination methods (bin-means averaging with five and ten bins, often called "delta plots", and linear transform pooling) applied to ten samples of size 60. Bin-means averages are plotted at the midpoints of their bins. The distributions are among those specified in Tables 1, 2, and 3. The top and bottom sets of plots have been translated upward by 55 ms and downward by 50 ms, respectively, to avoid overlap

8. Conclusions

The sample sizes of reaction-time data collected under constant conditions are inevitably small, too small to enable reliable estimation of the shape of the underlying distribution: Within subjects there are practice and fatigue effects, and across subjects there are individual differences. One possibility (shape invariance) is that these differences influence only the location and spread of RT distributions, but not their shapes. An alternative possibility (shape variation) is that the differences also influence their shapes. Given shape invariance, we would like to combine the samples in such a way as to estimate the shape that is shared by their parent distributions. Given shape variation, we would like to combine the samples in such a way as to estimate a shape that is some sort of average of the different shapes of their parent distributions.

This paper considers two combination methods that have been proposed and reported in the psychological literature: bin-means averaging with its associated bin-means histograms (sometimes called "Vincentizing"), and quantile averaging. And it introduces two additional methods: shape averaging (which depends on L-moments) and linear-transform pooling. Simulations in which the underlying distributions are specified are used to evaluate the estimates produced by the four methods. In some of the simulations, the parent distributions from which samples are drawn and combined have the same shape but differ in location and scale; in other simulations they differ in shape.

Most of the evaluations use shape measures based on L-moments, measures that were introduced by Hosking in 1990. However, the conclusions are also supported when the measures of distribution shape are the traditional ones, based on central moments. Appendices provide an introduction to L-moments, and demonstrate their superiority to central moments with respect to bias and outlier sensitivity.

The accuracies of the estimates produced by the four combination methods are found to differ greatly, and to favor the two new methods. Indeed, the estimates

produced by bin-means histograms are sufficiently biased so that published conclusions based on that method should probably be revisited. With five or ten bins, the bin-means averages, themselves, can be useful, but no more useful than linear-transform pooling. To determine an approximation to the shape of a distribution of observed RTs, or to select a theoretical distribution that is approximately consistent with it, we should use either linear-transform pooling, or shape averaging. R code to implement these methods is provided in Appendix H. To estimate shape differences, linear-transform pooling should be used.

Appendix A: definitions and properties of L-moments

The first six L-moments are defined as follows: Let $X_{m:n}$ be the m^{th} order-statistic in a sample of size n, and E denote expectation.[29]

$$\lambda_1 = E(X)$$
$$\lambda_2 = E(X_{2:2} - X_{1:2})/2$$
$$\lambda_3 = E(X_{3:3} - 2X_{2:3} + X_{1:3})/3$$
$$\lambda_4 = E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4})/4$$
$$\lambda_5 = E(X_{5:5} - 4X_{4:5} + 6X_{3:5} - 4_{2:5} + X_{1:5})/5$$
$$\lambda_6 = E(X_{6:6} - 5X_{5:6} + 10X_{4:6} - 10X_{3:6} + 5X_{2:6} - X_{1:6})/6$$

To understand these L-moments, it helps to rewrite the expressions inside the parentheses as follows:

For λ_3 ,

$$(X_{3:3} - X_{2:3}) - (X_{2:3} - X_{1:3}).$$

Thus, λ_3 will be zero if the distribution is symmetric, such that the mean separation between the middle and third order-statistic is equal to the that between the middle and first order-statistic; if it is larger (smaller), it will be positive (negative).

For λ_4 ,

$$(X_{4:4} - X_{1:4}) - 3(X_{3:4} - X_{2:4}).$$

If the means of the four order-statistics are equally spaced, the separation between the means of the two middle ones will be one third of the that between the first and fourth, and we will have $\lambda_4 = 0$. If the separation is less (more) than one third, e.g., if either or both tails are heavy (light) relative to the middle of the distribution, λ_4 will be positive (negative).[30]

For λ_5 ,

$$\left[(X_{5:5} - X_{4:5}) - (X_{2:5} - X_{1:5})\right] - 3\left[(X_{4:5} - X_{3:5}) - (X_{3:5} - X_{2:5})\right].$$

The value of λ_5 indicates how much of the asymmetry measured by λ_3 is due to the tails (large first term) rather than the middle region of the distribution (large second term).

For λ_6 ,

$$\left[(X_{6:6} - X_{1:6}) - 5(X_{4:6} - X_{3:6})\right] - 5\left[(X_{5:6} - X_{4:6}) + 5(X_{3:6} - X_{2:6})\right].$$

This will be large if either tail and/or the peak (first term) is large relative to the "shoulders" (second term) of the distribution. Thus, unlike λ_4, and supplementing it, λ_6 distinguishes the shoulders from other parts of the distribution.

For $k \geq 3$, the L-moments are usually standardized: $\lambda_{sk} = \lambda_k / \lambda_2$, which defines L-skewness (λ_{s3}) and L-kurtosis (λ_{s4}).

The standardized L-moments are constrained by several inequalities, including $|\lambda_{sk}| < 1$, $(k \geq 3)$; $\lambda_{s4} \geq -1/4$; $\lambda_{s6} \geq -1/6$; and $4\lambda_{s4} \geq 5(\lambda_{s3})^2 - 1$ (Hosking, 1996; Jones, 2004). Given the limited ranges of λ_{sk} when $k \geq 3$, small differences in their values matter greatly.

[29] The order statistics of a sample are the sample values in ascending order. One approach to computing the n^{th} L-moment, λ_n, of a sample would be to determine the mean of each of the m^{th} order statistics ($m = 1, 2,.., n$) of all possible combinations of n values of the sample, and then to combine these means as in the equation for λ_n below. See Wang (1996). For example, suppose a sample of size $n = 5$ with values 1, 2, 3, 4, and 5, and we want to calculate λ_3 for that sample. The set of all possible combinations of size 3 is: [(1,2,3), (1,2,4), (1,2,5), (1,3,4), (1,3,5), (1,4,5), (2,3,4), (2,3,5), (2,4,5), and (3,4,5)]. (These would appear with equal probability in a random sample.) The second order-statistics of this set are 2, 2, 2, 3, 3, 4, 3, 3, 4, and 4. Their mean, $E(X_{2:3}) = 3.0$. Similarly, $E(X_{1:3}) = 1.5$, and $E(X_{3:3}) = 4.5$. Thus, $\lambda_3 = [1.5 - (2 \times 3.0) + 4.5] = 0$, as expected, given that the sample, thought of as a distribution, is symmetric.

[30] Thus, unlike the measure of kurtosis, β_2, based on the standardized fourth central moment, the meaning of L-kurtosis is clear.

Appendix B: bias of shape measures based on L-moments and C-moments

Among the advantages of measures of distribution shape based on L-moments over those based on C-moments is that the former are less biased, especially for small samples. This difference is demonstrated by the findings pictured below (Fig. B1 and B2).

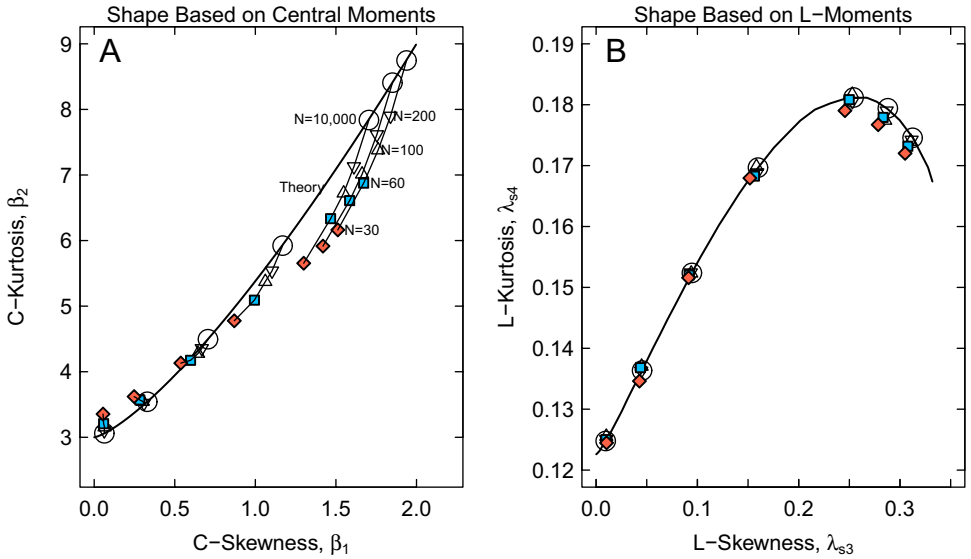

Fig. B1 Bias versus Sample Size of Two Pairs of Shape Measures for Seven Ex-Gaussian Distributions. Plotted values are based on means of 5000 replications. Estimates of the same shape based on different sample sizes are connected by line segments. Panel **A**: Shape measures β_1 and β_2, based on central moments. Panel B: Shape measures λ_{s3} and λ_{s4} based on L-moments. Centers of the large open circles represent the "true values" of these measures (based on samples of size 10,000). Panel **A** shows that even for sample sizes as large as 100 and 200 (triangles), the estimated shape values based on central moments are far from the true values, especially for more skewed distributions, whereas the corresponding estimates in Panel **B**, based on L-moments, have relatively little bias

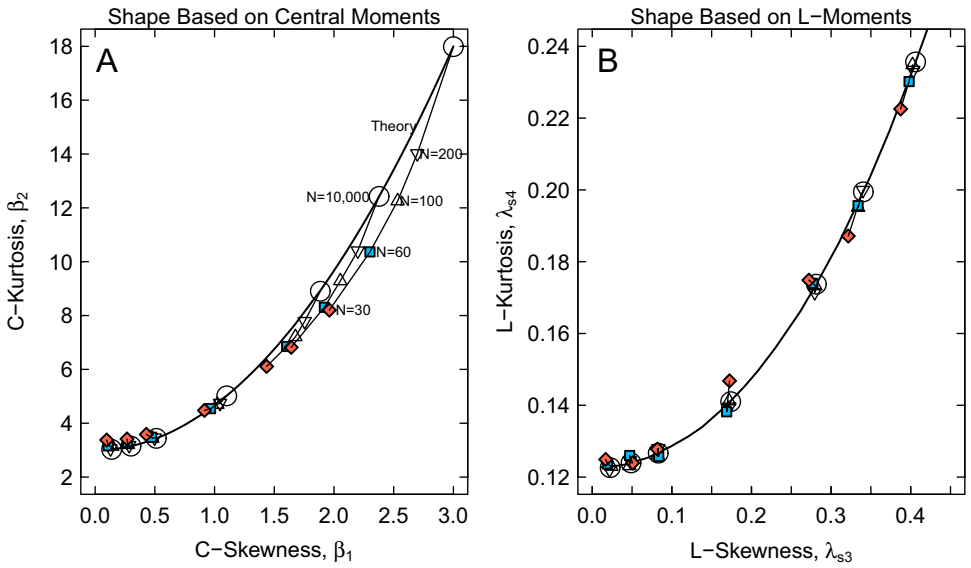

Fig. B2 Bias versus Sample Size of Two Pairs of Shape Measures for Seven Inverse-Gaussian (Wald) Distributions. Results are similar to those for the ex-Gaussian distributions in Fig. B1

Appendix C: outlier effects on shape measures based on L-moments and C-moments

Among the advantages of measures of distribution shape based on L-moments over those based on C-moments is that the former are less influenced by outliers. This difference is demonstrated by the findings pictured below, which show the effects on mean shape measures of introducing one high outlier among 200 RTs, for seven ex-Gaussian distributions (Fig. C1) and seven inverse-Gaussian (Wald) distributions (Fig. C2).

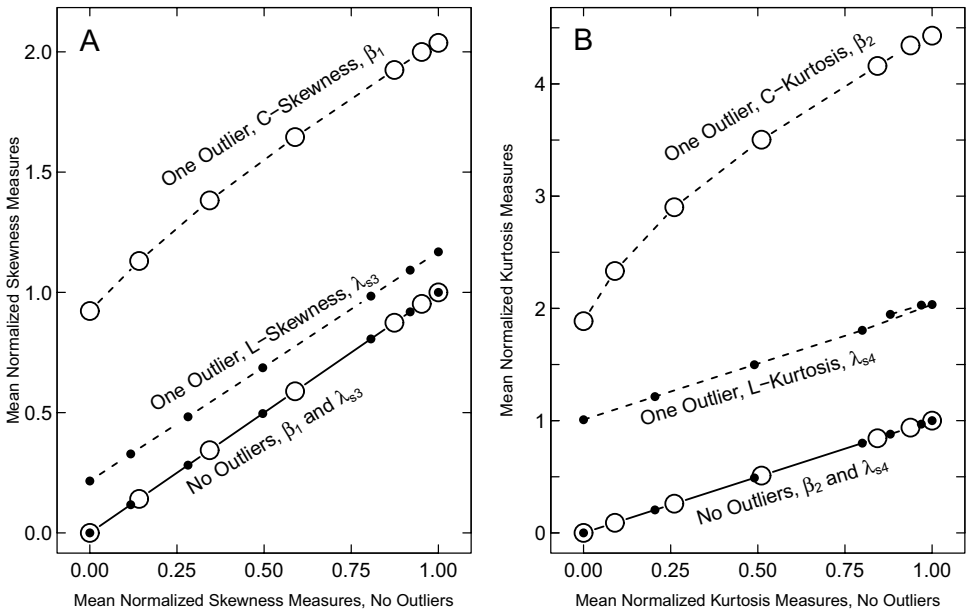

Fig. C1 Effects on mean shape measures of introducing one high outlier among 200 RTs, for seven ex-Gaussian distributions. Plotted values are based on means of 10,000 replications. For each distribution and each replication, the outlier was introduced by adding 3×SD to the largest observation among 200 sampled observations. Shape measures based on L-moments (λ_{s3} and λ_{s4}) and based on C-moments (β_1 and β_2) were determined for each sample with and without the outlier. Panel **A**: Means of the two skewness measures with and without the outlier, after linear normalization such that without the outlier, measures for seven different distributions increase linearly from zero to one. Small filled circles and large open circles represent values of λ_{s3} and β_1, respectively. Panel **B**: Means of the two kurtosis measures with and without the outlier, represented similarly

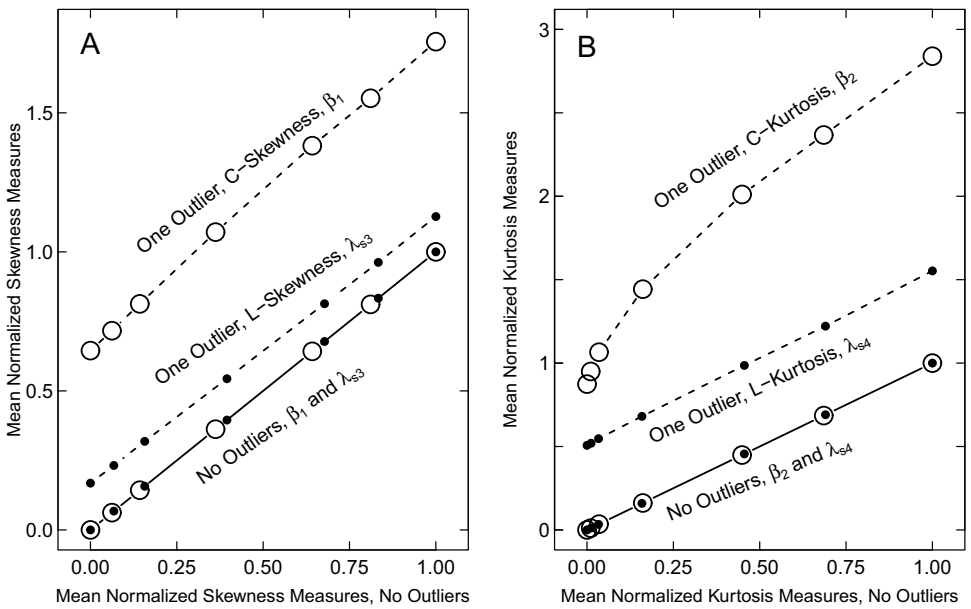

Fig. C2 The same analysis as in Fig. C1, but with samples drawn from seven inverse-Gaussian (Wald) distributions

Appendix D: density functions of nine basic distributions

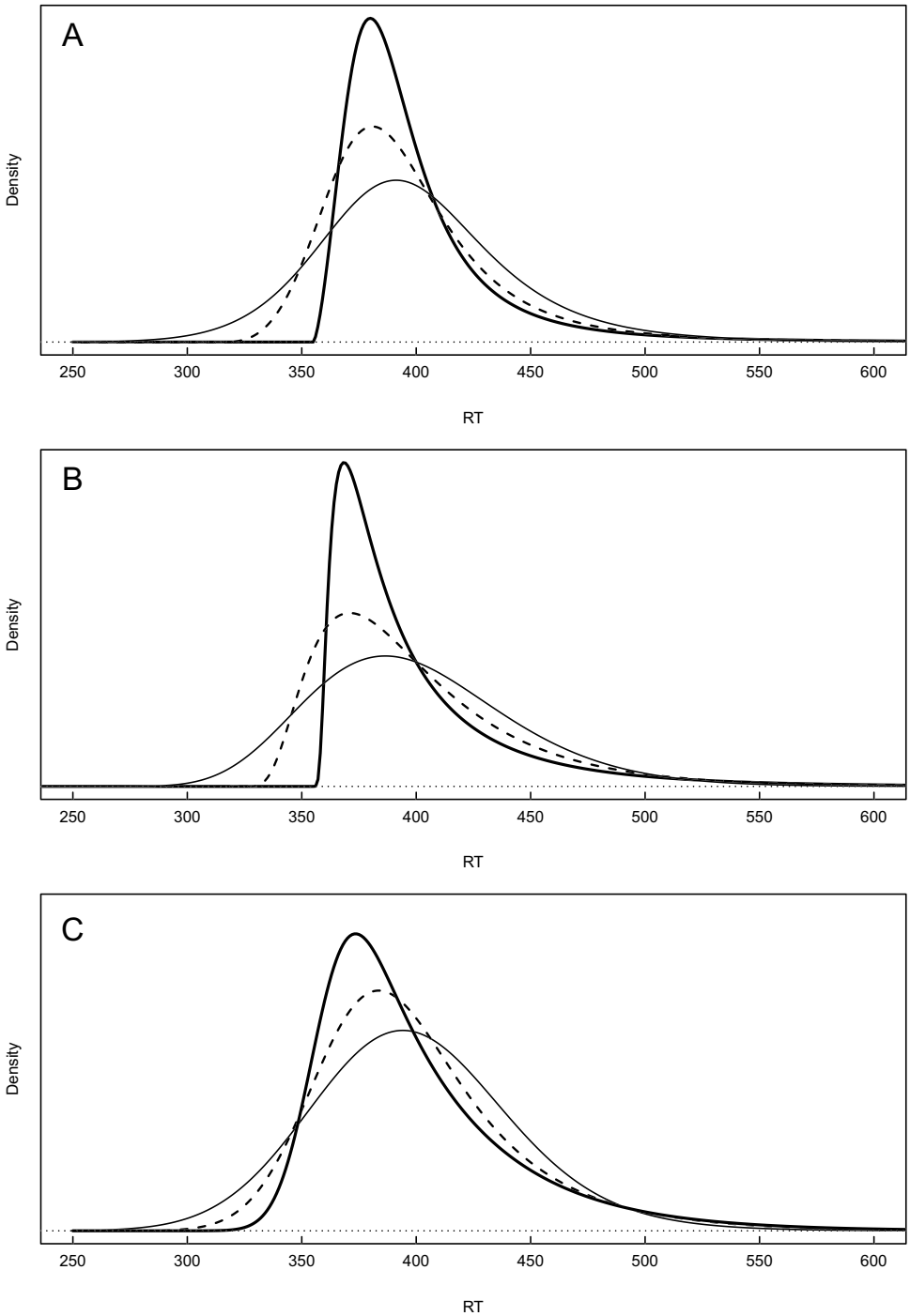

Fig. D1 Density Functions of the Nine Basic Distributions. Panel A: Generalized logistic distributions. From the distribution with highest to lowest peak, values of L-skewness are 0.400, 0.250, 0.100, and values of L-kurtosis are 0.300, 0.219, 0.175, respectively. Panel B: Shifted inverse Gaussian (Wald) distributions. From the distribution with highest to lowest peak, values of L-skewness are 0.400, 0.249, and 0.100, and values of L-kurtosis are 0.232, 0.162, and 0.129, respectively. Panel C: Ex-Gaussian distributions. From the distribution with highest to lowest peak, values of L-skewness are 0.264, 0.172, and 0.050, and values of L-kurtosis are 0.181, 0.172, and 0.138, respectively. All distributions were adjusted to have means of 400 units and standard deviations of 45 units

Appendix E: choice among variants of linear-transform pooling

Tables E1, E2, E3, E4, E5 and E6 show the mean estimation error (Tables E1A and E2A, etc.) and the mean bias (Tables E1B and E2B, etc.) for each of the thirty variants of the linear-transform pooling method, defined by combining five measures of location: the mean, the 20% winsorized mean (winmn), the 20% trimmed mean (trimmn), the mean of the 0.25 and 0.75 quantiles (quartmn), and the median, with six measures of scale: Qn (Rousseeuw & Croux, 1993), the second L-moment (lm2), the square root of the 20% winsorized variance (swv, Wilcox, 2017), the median absolute deviation (mad), the interquartile range (iqr), and the standard deviation (sdev). The values in each table are the means from two independent simulations, each involving 500 replications, and averaged over the nine basic distributions. Sample sizes of 60 and 100 were chosen because they seemed to approximate the sizes most likely to be used.

When ten samples from the same parent distributions were combined (Tables E1 and E2) the mean estimation error was smallest for the mean + Qn variant. When four samples from the same parent distributions were combined (Tables E3 and E4) the mean estimation error was smallest for the mean + lm2 variant. However, all six tables show that the mean bias associated with the mean + lm2 variant is considerably greater than with the mean + Qn variant, while their mean estimation errors do not differ greatly. I therefore selected the mean+Qn variant of the linear-transform method for comparison with the other methods.

It is worth noting, however, that relative to the differences in mean estimation error produced by the various methods (Fig. 2), the range of values of the mean estimation error

Table E1 Ten samples, same-shape parents, $N = 60$

A. 100 × Mean Estimation Error

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	3.687	3.863	3.888	3.914	3.961	4.424	3.956
winmn	3.737	3.888	3.955	4.013	4.042	4.509	4.024
trimmn	3.732	3.890	3.977	4.031	4.062	4.494	4.031
quartmn	3.763	3.917	3.976	4.040	4.052	4.549	4.050
median	3.751	3.999	4.040	4.063	4.122	4.611	4.098
Mean	3.734	3.911	3.967	4.012	4.048	4.517	4.032

B. 100 × Mean Bias

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	1.599	7.425	2.695	2.760	3.023	15.498	5.501
winmn	0.682	7.889	1.561	1.876	1.939	16.794	5.124
trimmn	0.609	7.729	1.796	2.099	2.165	16.150	5.091
quartmn	0.872	8.256	1.612	1.926	1.870	17.520	5.343
median	0.627	8.448	2.304	2.391	2.664	16.621	5.509
Mean	0.878	7.950	1.994	2.211	2.332	16.517	5.314

Table E2 Ten samples, same-shape parents, $N = 100$

A. 100 × Mean Estimation Error

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	2.836	2.915	2.918	2.940	2.962	3.286	2.976
winmn	2.868	2.930	2.947	2.991	2.997	3.337	3.012
trimmn	2.866	2.931	2.958	3.002	3.008	3.319	3.014
quartmn	2.876	2.945	2.957	3.003	3.002	3.362	3.024
median	2.873	2.984	2.990	3.017	3.039	3.373	3.046
Mean	2.864	2.941	2.954	2.991	3.002	3.336	3.015

B. 100 × Mean Bias

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	0.749	3.210	1.171	1.243	1.353	7.650	2.562 ‘
winmn	0.412	3.404	0.792	0.959	0.983	8.110	2.443
trimmn	0.385	3.340	0.853	1.038	1.051	7.727	2.399
quartmn	0.483	3.559	0.833	0.985	0.982	8.443	2.548
median	0.382	3.647	0.984	1.091	1.182	7.832	2.520
Mean	0.483	3.431	0.926	1.063	1.110	7.952	2.494

Table E3 Four samples, same-shape parents, $N = 60$

A. 100 × Mean Estimation Error

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	5.569	5.538	5.671	5.700	5.721	5.744	5.657
winmn	5.662	5.545	5.770	5.842	5.841	5.802	5.744
trimmn	5.666	5.554	5.785	5.864	5.856	5.803	5.755
quartmn	5.671	5.553	5.777	5.853	5.846	5.817	5.753
median	5.682	5.620	5.830	5.893	5.899	5.885	5.801
Mean	5.650	5.562	5.767	5.830	5.833	5.810	5.742

B. 100 × Mean Bias

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	1.571	5.835	2.271	2.320	2.502	11.689	4.364
winmn	0.782	6.127	1.403	1.627	1.677	12.562	4.030
trimmn	0.722	6.020	1.566	1.743	1.824	12.089	3.994
quartmn	0.939	6.377	1.459	1.699	1.626	13.038	4.189
median	0.778	6.572	1.939	1.915	2.181	12.438	4.304
Mean	0.958	6.186	1.728	1.860	1.962	12.364	4.176

Table E4 Four samples, same-shape parents, $N = 100$

A. 100 × Mean Estimation Error

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	4.394	4.348	4.445	4.445	4.470	4.479	4.430
winmn	4.455	4.352	4.506	4.524	4.537	4.516	4.482
trimmn	4.453	4.355	4.513	4.529	4.543	4.508	4.483
quartmn	4.459	4.358	4.510	4.531	4.539	4.528	4.488
median	4.458	4.391	4.536	4.540	4.567	4.547	4.506
Mean	4.444	4.361	4.502	4.514	4.531	4.515	4.478

B. 100 × Mean Bias

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	0.577	2.454	0.863	0.845	0.970	5.779	1.915
winmn	0.288	2.575	0.607	0.655	0.726	6.069	1.820
trimmn	0.254	2.544	0.645	0.655	0.754	5.788	1.773
quartmn	0.355	2.652	0.658	0.704	0.726	6.284	1.896
median	0.302	2.789	0.739	0.639	0.840	5.855	1.860
Mean	0.355	2.602	0.703	0.700	0.803	5.955	1.853

Table E5 Four samples, different-shape parents, $N = 60$

A. 100 × Mean Estimation Error

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	5.749	5.559	5.850	5.890	5.906	5.767	5.787
winmn	5.972	5.608	6.073	6.200	6.164	5.894	5.985
trimmn	5.995	5.637	6.102	6.252	6.200	5.934	6.020
quartmn	5.975	5.610	6.078	6.203	6.175	5.896	5.990
median	6.017	5.732	6.157	6.300	6.255	6.066	6.088
Mean	5.942	5.629	6.052	6.169	6.140	5.911	5.974

B. 100 × Mean Bias

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	0.648	1.670	0.993	1.148	1.127	2.714	1.383
winmn	1.025	1.721	1.332	1.693	1.522	2.890	1.697
trimmn	1.089	1.746	1.406	1.817	1.602	2.912	1.762
quartmn	1.056	1.767	1.358	1.682	1.541	2.921	1.721
median	1.130	1.945	1.502	1.886	1.678	3.094	1.872
Mean	0.990	1.770	1.318	1.645	1.494	2.906	1.687

produced by the thirty variants of the linear-transform pooling method is small. For example, the range when four samples are combined and $N = 100$ (Table E4) is (4.35, 4.55), a difference of 0.20, while the range of means across all methods for this case (Fig. 2B) is (4.45, 8.95), a difference of 4.50.

Appendix F: why are bin-means histograms so bad?

A hint as to the reasons for the large bias of the results of bin means averaging is provided by noting how the bin-means histogram represents the tails of a distribution. Recall that to represent a distribution, a set of m bins between $m + 1$ equally spaced proportions is defined, and the means $\{X_{k\bullet}\}$, $(k = 1, 2, \ldots, m)$, of the values in each bin determined. The bin-means histogram contains m equal-area rectangles bounded by the successive pairs of these means. As noted by van Zandt (2000, p. 430), such histograms exclude the data below the smallest and above the greatest of the m means. To demonstrate this, I created a distribution by combining a uniformly distributed center with low and high tails. A standard histogram of this distribution together with its density function are shown in Panel A of Fig. F1. The same data are represented as bin-means histograms with 6 and 15 bins, respectively, in Panels B and C of Fig. F1. Note how much area under the density function is outside the range of the bin-means histograms.

Table E6 Four samples, different-shape parents, $N = 100$

A. 100 × Mean Estimation Error

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	4.467	4.318	4.496	4.540	4.527	4.497	4.474
winmn	4.631	4.373	4.646	4.753	4.701	4.624	4.621
trimmn	4.647	4.393	4.670	4.789	4.728	4.650	4.646
quartmn	4.633	4.379	4.651	4.756	4.705	4.632	4.626
median	4.663	4.457	4.713	4.822	4.771	4.733	4.693
Mean	4.608	4.384	4.635	4.732	4.686	4.627	4.612

B. 100 × Mean Bias

	Qn	lm2	swv	mad	iqr	sdev	Mean
mean	0.677	1.171	0.761	0.879	0.849	1.993	1.055
winmn	0.868	1.273	0.939	1.203	1.072	2.186	1.257
trimmn	0.912	1.310	1.002	1.289	1.137	2.213	1.310
quartmn	0.885	1.304	0.958	1.202	1.084	2.214	1.274
median	0.930	1.444	1.081	1.340	1.201	2.330	1.388
Mean	0.854	1.300	0.948	1.183	1.069	2.187	1.257

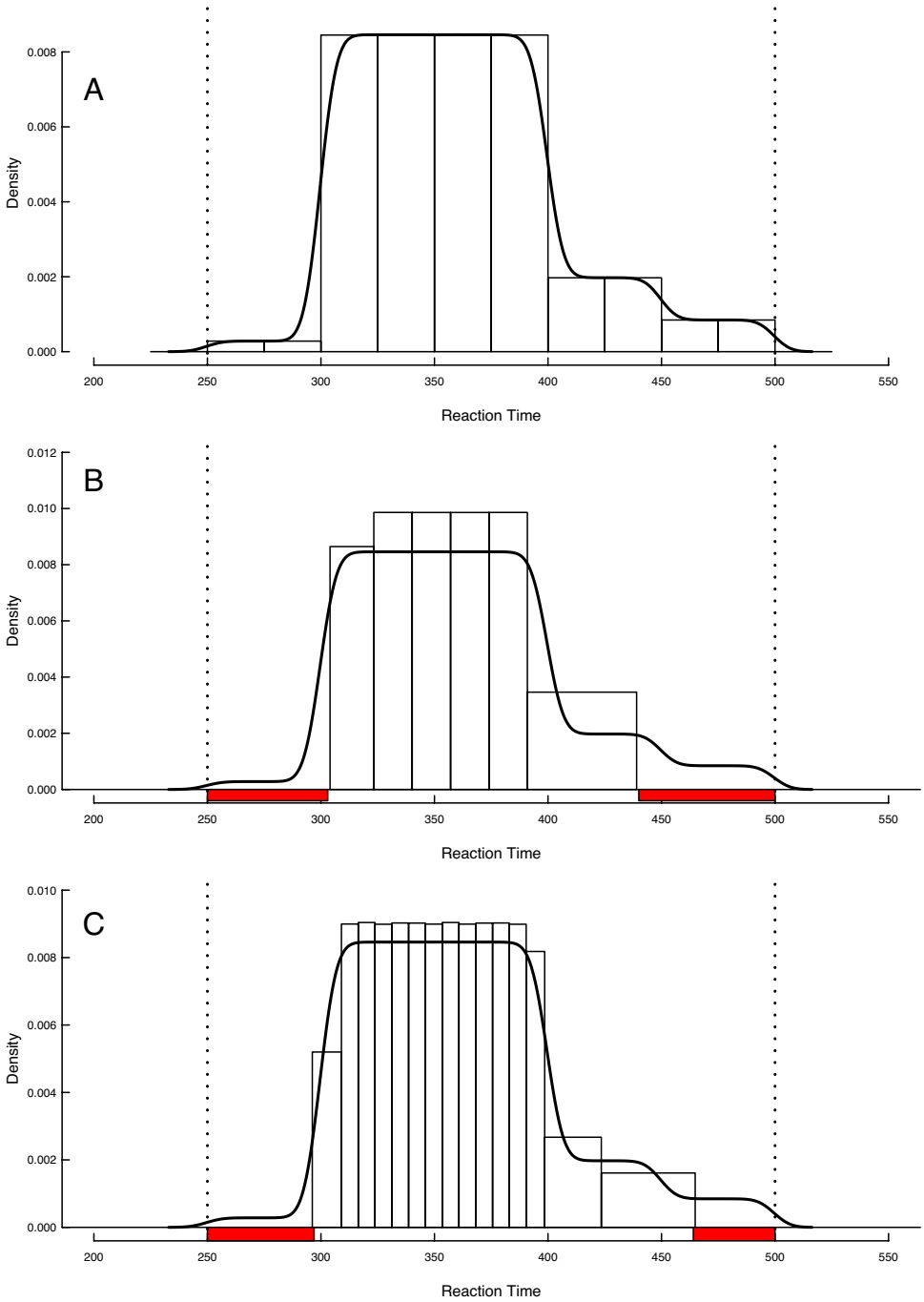

Fig. F1 Standard and Bin−Means Histograms. Representations of a distribution by standard and bin-means histograms. Panel **A**. Standard histogram of a distribution composed of a central uniformly distributed portion together with low and high tails. Also shown is a density function for the distribution. Panel **B**. Bin-means histogram of the same data with 6 bins. The same density function is shown. The red regions under the x-axis indicate values in the distribution that are excluded from the binmeans histogram. Panel **C**. Bin-means histogram of the same data with 15 bins. The same density function is shown. The red regions under the x-axis indicate values in the distribution that are excluded from the bin-means histogram

The L-moments associated with the sample values and with the two bin-means histograms are shown in Table F1. Consistent with the bias shown in Figs. 4, 5, and 6, we see that the values associated with the bin-means histograms are underestimates.

Table F1 L-moments of sample and bin-means histogram representations

Source	λ_{s3}	λ_{s4}	λ_{s5}	λ_{s6}
Sample	0.127	0.119	0.057	0.052
6 Bins	0.087	0.068	0.031	0.011
15 Bins	0.102	0.089	0.050	0.035

Appendix G: R code for selected functions

The following function starts with a sample, and generates bin means ("Vincentiles") with nbin bins.

```
vincentize <- function(data,nbins){
datalen <- length(data)
data.aug <- sort(rep(data,nbins))
data.augmat <- matrix(data.aug,nrow=datalen,ncol=nbins)
binmeans <- apply(data.augmat,2,mean)
return(binmeans)
}
```

quantiles were generated using the type-8 quantile function:

```
quantile(sample, prob=seq(0,1,.01), type=8).
```

Given averaged quantiles, a sample was generated using the following function (a deterministic version of inverse transform sampling), which distributes nobs observations uniformly in each interval, thus creating total obsns = nobs*(length(quantiles)-1).

The same function was used to generate a sample from bin-means averaging, replacing quantiles by bin means.

```
distrib.from.quantiles <- function(nobs,quantiles){
len <- length(quantiles)
interval.minima <- quantiles[1:(len-1)]
interval.maxima <- quantiles[2:len]
values <- 999999
for(kinterval in 1:(len-2)){
new.values <- seq(interval.minima[kinterval],interval.maxima[kinterval],
        length.out=nobs)
#remove largest, which will be repeated as first among next set of new.values
#for all except last set of new values
new.values.adj <- new.values[1:(nobs-1)]
values <- c(values,new.values.adj)
}
kinterval <- len-1
last.new.values <- seq(interval.minima[kinterval],interval.maxima[kinterval],
        length.out=nobs)
#for last interval, last new value is not removed
values <- c(values,last.new.values)
values <- values[-1]
return(values)
}
```

Appendix H: R code to implement linear-transform pooling and shape averaging

Each sample of reaction-time data is a vector. The input to the program, called "sample.list" is a list of these vectors. The program outputs are "pool" and "meanshape".

```
##############################################
#Generate data for this toy example
samp1 <- c(428,480,508,388,485,488,450,451,450,439,453,435,443,472)
samp2 <- c(444,417,407,397,399,480,423,408,409,431,403,384,434,394,420,397,448)
samp3 <- c(436,429,435,401,415,385,410,441,436,493,418,424,442,420,421,417,412,434,435)
sample.list <- list(samp1,samp2,samp3)
len <- length(sample.list)
##############################################
#required packages
library(lmom)
library(robustbase)
##############################################
######  Linear-Transform Pooling  ############
##############################################
#define adjustment function
adjust <- function(sample,targetlocation,targetscale){
xadj <- targetlocation + (sample - samplelocation)*(targetscale/samplescale)
return(xadj)
}
##############################################
#Use mean location and mean scale as targets (targets are arbitrary)
samplocations <- rep(NA,len)
sampscales <- rep(NA,len)
for(ksamp in 1:len){
sample <- sample.list[[ksamp]]
samplocations[ksamp] <- mean(sample)
sampscales[ksamp] <- Qn(sample)
}
targetlocation <- round(mean(samplocations))
targetscale <- round(mean(sampscales))
##############################################
#adjust samples and combine them
adjusted.list <- vector("list",len)
for(ksamp in 1:len){
sample <- sample.list[[ksamp]]
samplelocation <- mean(sample)
samplescale <- Qn(sample)
adjusted.list[[ksamp]] <- adjust(sample,targetlocation=targetlocation,
        targetscale=targetscale)
}
pool <- unlist(adjusted.list)
##############################################
##############################################
######       Shape Averaging           ######
##############################################
#define function to compute shape values
get.shape <- function(x){
values <- samlmu(x,nmom=6)[3:6]
names(values) <- c("lskew","lkurt","lmom.s5","lmom.s6")
return(values)
}
##############################################
shape.mat <- matrix(nrow=len,ncol=4)
rnames <- paste("samp.",1:len,sep="")
cnames <- c("lskew","lkurt","lmom.s5","lmom.s6")
dimnames(shape.mat) <- list(rnames,cnames)
for(ksamp in 1:len){
sample <- sample.list[[ksamp]]
shape.mat[ksamp,] <- get.shape(sample)
}
mean.shape <- apply(shape.mat,2,mean)
##############################################
##############################################
#it may be of interest to compare shape values for the pooled adjusted samples
#using get.shape(pool) to the values in shape.mat and to mean.shape.
R%%>shape.mat
           lskew   lkurt   lmom.s5 lmom.s6
samp.1 -0.013496 0.21974 -0.177904 0.14647
samp.2  0.236601 0.16091  0.061505 0.13365
samp.3  0.094148 0.31192  0.138105 0.34671
R%%>mean.shape
    lskew     lkurt   lmom.s5   lmom.s6
0.1057511 0.2308541 0.0072356 0.2089434
R%%>get.shape(pool)
   lskew    lkurt  lmom.s5  lmom.s6
0.103178 0.221170 0.034115 0.171407
```

Acknowledgements I thank Vincent Hurtubise and Mark Davidson for computer support.

Code availability R code for the simulations will be made available by the author to anyone who plans to pursue these investigations further.

References

An, H., Zhang, K., Oja, H., & Marron, J. (2023). Variable screening based on Gaussian Centered L-moments. *Computational Statistics and Data Analysis, 179*, 107632.

Anders, R., Alario, F.-X., & Van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychological Methods, 21*, 309–327.

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 514–544.

Asquith, W. H. (2007). L-moments and TL-moments of the generalized lambda distribution. *Computational Statistics & Data Analysis, 51*, 4484–4496.

Asquith, W. H. (2011). *Distributional analysis with L-moment statistics using the R environment for statistical computing*. Create Space Independent Publishing Platform.

Asquith, W. H. (2014). Parameter estimation for the 4-parameter asymmetric exponential power distribution by the method of L-moments using R. *Computational Statistics & Data Analysis, 71*, 955–970.

Asquith, W. H. (2021). lmomco---L-moments, censored L-moments, trimmed L-moments, L-comoments, and many distributions. R package version 2.3.7, Texas Tech University

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Current Directions in Psychological Science, 20*, 160–166.

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analysis of semantic priming. *Journal of Memory and Language, 59*, 495–523.

Christie, L. S., & Luce, R. D. (1956). Decision structure and time relations in simple choice behavior. *Bulletin of Mathematical Biophysics, 18*, 89–112.

Colonius, H., & Vorberg, D. (1994). Distribution inequalities for parallel models with unlimited capacity. *Journal of Mathematical Psychology, 38*, 35–58.

Cousineau, D., Thivierge, J.-P., Harding, B., & Lacouture, Y. (2016). Constructing a group distribution from individual distributions. *Canadian Journal of Experimental Psychology, 70*, 253–277.

De Jong, R., Liang, C.-C., & Lauber, E. (1994). Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus-response correspondence. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 731–750.

Dawson, M. R. W. (1988). Fitting the ex-Gaussian equation to reaction time distributions. *Behavior Research Methods, Instruments, & Computers, 20*, 54–57.

Ellinghaus, R., & Miller, J. (2018). Delta plots with negative-going slopes as a potential marker of decreasing response activation in masked semantic priming. *Psychological Research Psychologische Forschung, 82*, 590–599.

Gondan, M., & Vorberg, D. (2021). Testing trisensory interactions. *Journal of Mathematical Psychology, 101*, 102513. https://doi.org/10.1016/j.jmp.2021.102513

Headrick, T. C. (2011). A characterization of power method transformations through L-Moments. *Journal of Probability and Statistics, 2011*, Article ID 497463.

Heathcote, A. (1996). RTSYS: A DOS application for the analysis of reaction time data. *Behavior Research Methods, Instruments, & Computers, 28*, 427–445.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin, 109*, 340–347.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology, 69*, 382–386.

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society B, 52*, 105–124.

Hosking, J. R. M. (1996). *Some theoretical results concerning L-moments*. Research Report RC 14492 (revised version) IBM Research Division, Yorktown Heights, NY.

Hosking, J. R. M. (1992). Moments or L-moments? An example comparing two measures of distributional shape. *The American Statistician, 16*, 186–189.

Hosking, J. R. M. (2006). On the characterization of distributions by their L-moments. *Journal of Statistical Planning and Inference, 136*, 193–198.

Hosking, J. R. M. (2007). Some theory and practical uses of trimmed L-moments. *Journal of Statistical Planning and Inference, 43*, 299–314.

Hosking, J. R. M. (2019). L-Moments. R package, version 2.8. https://CRAN.R-project.org/package=lmom. Accessed December 2022.

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional frequency analysis: An approach based on L-moments*. Cambridge University Press.

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *American Statistician, 50*, 361–365. https://doi.org/10.2307/2684934

Jiang, Y., Rouder, J. N., & Speckman, P. L. (2004). A note on the sampling properties of the Vincentizing (quantile averaging) procedure. *Journal of Mathematical Psychology, 48*, 186–195.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions* (Vol. 1). Wiley.

Johnson, R. L., Staub, A., & Fleri, A. M. (2012). Distributional analysis of the transposed-letter neighborhood effect on naming latency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1773–1779.

Jones, M. C. (2004). On some expressions for variance, covariance, skewness and L-moments. *Journal of Statistical Planning and Inference, 126*, 97–106.

Karvanen, J., & Nuutinen, A. (2008). Characterizing the generalized lambda distribution by L-moments. *Computational Statistics & Data Analysis, 52*, 1971–1983.

Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics* (Vol. 1, 3rd ed.). New York: Hafner.

Kjeldsen, T. R., Ahn, H., & Prosdocimi, I. (2017). On the use of a four-parameter kappa distribution in regional frequency analysis. *Hydrological Sciences Journal, 62*, 1354–1363.

Little, D. R., Altieri, N., Fific, M., & Yang, C.-T. (Eds.) (2017). *Systems factorial technology.* Academic.

Luce, R. D. (1986). *Response times*. Oxford University Press.

Lombardi, L., D'Alessandro, M., & Colonius, H. (2019). A new nonparametric test for the race model inequality. *Behavior Research Methods, 51*, 2290–2301.

Mackenzie, I. G., Mittelstadt, V., Ulrich, R., & Leuthold, H. (2022). The role of temporal order of relevant and irrelevant dimensions within conflict tasks. *Journal of Experimental Psychology: Human Perception and Performance.* Advance online publication. https://doi.org/10.1037/xhp0001032

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L., & di Palma, M. A. (2022). *robustbase: Basic Robust Statistics.* R package version 0.95-0. http://CRAN.R-project.org/package=robustbase. Accessed December 2022.

Marzuki, M., Randeu, W. L., Kozu, T., Shimomai, T., Schonhuber, M., & Hashiguchi, H. (2012). Estimation of raindrop size distribution parameters by maximum likelihood and L-moment methods: Effect of discretization. *Atmospheric Research, 112*, 1–11.

McGill, W. J. (1963). Stochastic latency mechanisms. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 309–360). Wiley.

Miller, J. O. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14*, 247–279.

Miller, J., & Schwarz, W. (2021). Delta plots for conflict tasks: An activation-suppression race model. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-021-01900-5

Miller, R., Scherbaum, S., Heck, D. W., Goschke, T., & Enge, S. (2017). On the relation between the (censored) shifted Wald and the wiener distribution as measurement models for choice response times. *Applied Psychological Measurement, 42*, 116–135.

Mittelstädt, V., Miller, J., Leuthold, H., Mackenzie, I. G., & Ulrich, R. (2022). The time-course of distractor-based activation modulates effects of speed–accuracy tradeoffs in conflict tasks. *Psychonomic Bulletin & Review, 29*, 837–854.

Navruz, G., & Ozdemir, F. (2020). A new quantile estimator with weights based on a subsampling approach. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12198

Pearson, E. S. (1963). Some problems arising in approximating to probability distributions, using moments. *Biometrika, 50,* 95–112. Possibly

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/. Accessed December 2022.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin, 86*, 446–461.

Ratcliff, R., & Smith, P. L. (2010). Perceptual discrimination in static and dynamic noise: The temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology. General, 139*, 70–94.

Reynolds, A., & Miller, J. (2009). Display size effects in visual search: Analyses of reaction time distributions as mixtures. *The Quarterly Journal of Experimental Psychology, 62*, 988–1009.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience: A silver jubilee* (pp. 611–653). M.I.T. Press.

Rouder, J. N., & Speckman, P. L. (2004). An evaluation of the Vincentizing method of forming group-level response time distributions. *Psychonomic Bulletin & Review, 11*, 419–427.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association, 88*, 1273–1283.

Royston, P. (1992). Which measures of skewness and kurtosis are best? *Statistics in Medicine, 11*, 333–343.

Schweickert, R., & Giorgini, M. (1999). Response time distributions: Some simple effects of factors selectively influencing mental processes. *Psychonomic Bulletin & Review, 6*, 269–288.

Schweickert, R., Fisher, D. L., & Sung, K. (2012). *Discovering cognitive architecture by selectively influencing mental processes*. World Scientific.

Schwarz, W., & Miller, J. (2012). Response time models of delta plots with negative-going slopes. *Psychonomic Bulletin & Review, 19*, 555–574.

Selaman, O. S., Said, S., & Putuhena, F. J. (2007). Flood frequency analysis for Sarawak using Weibull, Gringorten and L-Moments Formula. *Journal of The Institution of Engineers, Malaysia, 68*, 43–52.

Snodgrass, J. G., Luce, R. D., & Galanter, E. (1967). Some experiments on simple and choice reaction time. *Journal of Experimental Psychology, 75*, 1–17.

Spieler, D. H., & Balota, D. A. (1997). Bringing computational models of word naming down to the item level. *Psychological Science, 8*, 411–416.

Stasinopoulos, M. & Rigby, R. (2022). gamlss.dist: Distributions for Generalized Additive Models for Location Scale and Shape. R package version 6.0-3. https://CRAN.R-project.org/package=gamlss.dist. Accessed December 2022.

Sternberg, S. (1964). Estimating the distribution of additive reaction-time components. Paper presented at the meeting of the Psychometric Society, Niagara Falls, Canada, October.

Sternberg, S. (1973). Evidence against self-terminating memory search from properties of RT distributions. Paper presented at a meeting of the Psychonomic Society, St. Louis, November.

Sternberg, S. (2016). In defence of high-speed memory scanning. *The Quarterly Journal of Experimental Psychology, 69*, 2020–2075.

Sternberg, S., & Backus, B. T. (2015). Sequential processes and the shapes of reaction time distributions. *Psychological Review, 122*, 830–837.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25*, 251–260.

Thomas, E. A. C., & Ross, B. H. (1980). On appropriate procedures for combining probability distributions within the same family. *Journal of Mathematical Psychology, 21*, 136–152.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge University Press.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology, 39*, 321–359.

Ulrych, T. J., Velis, D. R., Woodbury, A. D., & Sacchi, M. D. (2000). L-moments and C-moments. *Stochastic Environmental Research and Risk Assessment, 14*, 50–68.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review, 7*, 424–465.

Vickers, D. (1979). *Decision processes in visual perception*. Academic.

Vogel, R. M., & Fennessey, N. M. (1993). L moment diagrams should replace product moment diagrams. *Water Resources Research, 29*, 1745–1752.

Vorberg, D. (1981). Reaction time distributions predicted by serial, self-terminating models of memory search. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology. Proceedings of the symposium in applied mathematics of the american mathematical society and the society for industrial and applied mathematics* (pp. 301–318). American Mathematical Society.

Wang, Q. J. (1996). Direct sample estimators of L moments. *Water Resources Research, 32*, 3617–3619.

Westfall, P. H. (2014). Kurtosis as Peakedness, 1905–2014. R.I.P. *The American Statistician, 68*, 191–195.

Wheeler, B. (2022). SuppDists: Supplementary Distributions. R package version 1.1-9.7. https://CRAN.R-project.org/package=SuppDists. Accessed December 2022.

Wilcox, R. R. (2017) Introduction to robust estimation and hypothesis testing (4th ed.). New York: Academic Press.

Withers, C. S., & Nadarajah, S. (2011). Bias-reduced estimates for skewness, kurtosis, L-skewness and L-kurtosis. *Journal of Statistical Planning and Inference, 141*, 3839–3861.

Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010). Reaction time distributions constrain models of visual search. *Vision Research, 50*, 1304–1311.

Yantis, S., Meyer, D. E., & Smith, J. E. K. (1991). Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin, 110*, 350–374.