Running head: BOOTSTRAP TEST OF DISTRIBUTION SHAPE

A Bootstrap Test of Shape Invariance Across Distributions

Jeffrey N. Rouder, Paul L. Speckman, Douglas Steinley, Michael S. Pratte and Richard D. Morey

University of Missouri-Columbia

Jeffrey N. Rouder

210 McAlester Hall

University of Missouri

Columbia, MO 65211

rouderj@missouri.edu

Abstract

The shape of a response time distribution provides valuable clues about the underlying mental processing. If a manipulation affects the shape of an RT distribution, then it is reasonable to suspect that the manipulation has done more than simply speed or slow the rate of processing. We develop a nonparametric bootstrap test of shape invariance. Simulations reveal the test is sufficiently powered to detect small shape changes in reasonably sized experiments while maintaining appropriate Type I error control. The test is simple and can be applied broadly in cognitive psychology. An application to a number priming experiment provides a demonstration of how shape changes may be detected.

A Bootstrap Test of Shape Invariance Across Distributions

Response time (RT), the time taken to complete a task, is a common dependent variable that has been used to draw inferences about the nature of mental processing (e.g., Luce, 1986). Although many researhcers draw conclusions from analysis of mean RT, it is increasingly common to see more sophisticated analyses of entire RT distributions. These more sophistcated analyses often provide additional constraint on cognitive and perceptual theories. Selective examples include Ashby, Tien, and Balakrishnan (1993); Dzhafarov (1992); Hockley (1984); Logan (1992); Ratcliff (1978); Ratcliff and Rouder (1998, 2000); Rouder (2000); Rouder, Ratcliff, and McKoon (2000); Spieler, Faust, and Balota, (1996); Theeuwes (1992, 1994); Townsend & Nozawa (1995); Van Zandt, Colonius, & Proctor (2000); and Vickers (1980).

Although there are several methods of analyzing distributions, we advocate that researchers consider how properties of *location*, *scale*, and *shape* change across conditions or populations (Rouder, Sun, Speckman, Lu, & Zhou, 2003; Rouder, Lu, Speckman, Sun, & Jiang, 2005). Figure 1 provides an example of these properties. The left panel shows the case when only location differs between distributions; the center and right panels show the same for scale and shape, respectively.

Location and scale are formally defined as follows: Let the density of a continuous random variable exist everywhere and be expressed as $f(t \mid \theta_1, \ldots, \theta_p)$, where $\theta_1, \ldots, \theta_p$ are parameters. Let $z = (t - \theta_1)/\theta_2$. We refer to the density as being in location-scale form if there exists a density g such that

$$f(t \mid \theta_1, \ldots, \theta_p) = \frac{1}{\theta_2} g\left(\frac{t - \theta_1}{\theta_2} \mid \theta_3, \ldots, \theta_p\right). \quad (1)$$

If Equation 1 holds, then θ_1 is referred to as a location parameter and θ_2 as a scale parameter. We will say that θ_3 through θ_p are shape parameters. Many random variables

have densities that can be expressed in location-scale form. For the normal, for example, $\theta_1 = \mu$, $\theta_2 = \sigma$, and $g(z) = (2\pi)^{-1}e^{-z^2/2}$. The location parameter corresponds to the mean, the scale parameter corresponds to the standard deviation, and there are no shape parameters. For the exponential with $f(x \mid \tau) = \tau^{-1}e^{-x/\tau}$, $x > 0$, there is no location parameter ($\theta_1 = 0$), $\theta_2 = \tau$, and $g(z) = e^{-z}$. The scale parameter corresponds to τ, and there are no shape or location parameters.

There is an asymmetric relationship between location, scale, and shape parameters and the central moments of a distribution. Changes in location certainly imply changes in mean but not in central moments of higher order than the mean. Changes in scale imply changes in even central moments including the variance. Changes in scale, in general, may imply changes in the mean too. For example, the exponential has a single parameter, the rate, whose inverse is a scale parameter. Increasing the rate not only decreases the variance, but decreases the mean as well. Changes in shape, in general, may imply changes in all moments.

We have advocated the following processing interpretation of these properties. Shape is the most important property as it may reflect the underlying mental architecture. Shape changes may be associated with changes in architecture, such as changes in mixtures, adding latent stages, or switching algorithms (e.g., from serial to parallel processing). If shape does change across manipulations, then documenting, exploring, and explaining these changes is surely a good route to better psychological theory. If shape is invariant across manipulations, then this invariance licenses the analysis of scale. Moreover, with shape invariance and stochastic dominance (see the center panel of Figure 1), the scale serves as an index of processing speed. Scale is not interpretable without shape invariance. For instance, it makes little sense to compare the scale of a normal distribution to that of an exponential distribution. Finally, shift indexes peripheral processes such as traducing the stimulus and executing motor commands (c.f.; Dhzafarov,

1992; Hsu, 1999; Ratcliff, 1978).

The above interpretation of location, scale, and shape places priority on shape. Therefore, it is important to develop methods of assessing shape invariance while allowing location and scale to vary. There are two approaches to the problem: the first is to specify a parametric form with an explicit shape parameter. Examples include the Weibull, log-normal, and inverse Gaussian. We show in the next section that the parametric approach is not robust to misspecification, and therefore, is not appropriate. The second approach is nonparametric. The nonparametric approach would be straightforward if the analyst knew the mean and variance of distributions to infinite precision. In this case, the distributions could be shifted and scaled until they had a mean of 0.0 and a variance of 1.0. Then, any shape differences would be manifest as differences in the shifted-scaled distributions and two-sample tests of distributional equality are seemingly appropriate. Unfortunately, the analyst only knows sample means and variances; therefore, testing for shape invariance is not equivalent to testing for distribution equality. In this paper, we develop a nonparametric sampling test of shape invariance.

Lack of Robustness of a Parametric Test

Before developing the nonparametric test, we highlight the problems of a parametric approach by exploring the robustness of a shape test that assumes a Weibull parametric form. The density of a three-parameter Weibull distribution may be given as

$$f(t;\psi,\theta,\beta) = \frac{\beta}{\theta}\left(\frac{t-\psi}{\theta}\right)^{\beta-1}\exp\left(-\left[\frac{t-\psi}{\theta}\right]^{\beta}\right),\ t>\psi,\ \theta,\beta>0.$$

In this parameterization parameters ψ, θ, and β serve as the location, scale, and shape of the distribution, respectively. To test the robustness of a Weibull parametric shape test, we performed two simulated experiment with each consisting of 50 participants who each observed 100 trials. In the first experiment, each participant's data came from a Weibull distribution with a location, scale, and shape parameter values of .262 s, .199 s, and 1.7,

respectively. Because the test assumes Weibull parametric forms, the test is well-specified for this experiment. In the second experiment, each participant's data came from an inverse-Gaussian distribution. The density of a three-parameter inverse-Gaussian distribution may be given as

$$f(x|\psi,\theta,\beta) = \sqrt{\frac{\theta}{2\pi}}(x-\psi)^{-3/2}\exp\left(\frac{-((x-\psi)\beta-\theta)^2}{2\theta(x-\psi)}\right), \quad x>\psi \;\; \theta,\beta>0.$$

In this parameterization parameters ψ, θ, and β serve as the location, scale, and shape of the distribution, respectively. Each participant's data had inverse-Gaussian location, scale, and shape parameters of .2 s, 1.2s, and 5, respectively. Because the test assumes the Weibull form, it is misspecified for these data. The contrast between the two experiments allows for study of the effects of misspecification. The solid line and dotted line in the left panel of Figure 2 show the densities of an inverse-Gaussian and a Weibull distribution, respectively, with the parameter values given above. These distributions are highly similar indicating that there is only a small degree of misspecification. Hence, we would hope that the Weibull parametric shape test yields reasonable results, even when applied to inverse-Gaussian data.

We fit two Weibull models. The general model consists of individualized location, scale, and shape parameters. Across the 50 participants, therefore, there are $3 \times 50 = 150$ parameters. The restricted model specified that each participant had the same shape; hence the restricted model had 50 shift parameters, 50 scale parameters, and a single shape parameter (total of 101 parameters). Whereas "true" shapes did not vary across participants, a well-calibrated test should yield a rate of Type I errors close to the nominal value. We estimated the Weibull models with maximum likelihood and assessed the validity of shape invariance with a likelihood ratio test (G^2; Reed & Cressie, 1988; Wilks, 1938). Details of maximization are as follows. For the general model, the negative log-likelihood was minimized for each individual separately. For each individual, there are

three parameters and the three-parameter minimization was performed with the Nedler-Mead Simplex algorithm (Nedler & Mead, 1967) in **R** with the `optim()` command. The case is a bit more complicated for the restricted model. We used a nested optimization design. In the inner loop, we minimized location and scale for each individual separately for a fixed common shape with the simplex algorithm. Then, in the outer loop, we found the best common shape that minimized negative log-likelihood with the algorithm from Brent (1973) as implemented in **R** with the `optimize()` command. Because each minimization call is done with respect to a small number of parameters, minimization was quick and reliable.

The Weibull model is not regular, and therefore the asymptotic distribution of the test statistic is not guaranteed to follow the chi-square distribution with 49 degrees-of-freedom ($df_{\text{general}} - df_{\text{restricted}} = 150 - 101 = 49$). To assess the effects of irregularity, we simulated the first experiment (Weibull-generated data) 300 times. The cumulative distribution function of the sampling distribution of G^2 across these 300 replicates is shown as line "W" in the right panel of Figure 2. The dotted line shows the theoretical chi-square distribution with 49 degrees-of-freedom. As can be seen, the observed distribution is fairly close to the chi-square distribution even though regularity is violated. The vertical line shows the criterial value for a nominal Type I error rate of .05. The observed type I error rate for this nominal value is .083, which represents only a modest inflation.

To assess how a seemingly small misspecification of parametric form affects the true sampling distribution of G^2, we performed 300 replications of the experiment with data generated from the inverse-Gaussian. Line "IG" shows the cumulative distribution function of the sampling distribution of G^2. Most of the distribution (85% of the 300 simulation experiments) is above the nominal $\alpha = .05$ criterial value indicating a massive Type I error inflation. Even though the data were generated with shape invariance, the

parametric Weibull test is useless as even small misspecification drive an unreasonably high Type I error rate. It is this uselessness that motivates the need for a nonparametric test.

Development of a Nonparametric Shape Test

We first develop a nonparametric shape test for the case for a single participant who provides data in each of two conditions. Let $\mathbf{x} = x_1, \ldots, x_J$ and $\mathbf{y} = y_1, \ldots, y_J$ denote vectors of J observations in the first and second conditions, respectively. Assume that the data in each condition are independent and identically distributed; e.g., $x_j \overset{iid}{\sim} X$ and $y_j \overset{iid}{\sim} Y$, $j = 1, \ldots, J$. The main question is whether distributions X and Y have the same shape; i.e., do $(X - \mu_X)/\sigma_X$ and $(Y - \mu_Y)/\sigma_Y$ have the same distribution. The first stage of the test is the construction of a statistic for shape change; the second stage is the estimation of the sampling distribution of the statistic through resampling.

Step #1: Transform the data. If one knew the population means and standard deviations for distributions X and Y, one could form the standardized samples, say

$$x_j^* = \frac{x_j - \mu_X}{\sigma_X}, \quad y_j^* = \frac{y_j - \mu_Y}{\sigma_Y}, \quad j = 1, \ldots, J, \tag{2}$$

and perform a nonparametric test for equality of two distributions such as the Kolmogorov-Smirnov two-sample test (Conover, 1971) or the Cramer-von Mises two-sample test (Anderson, 1962). Since the population means and standard deviations are unknown, it is natural to use the sample means and standard deviations estimated from the data, $\bar{x}$, $\bar{y}$, $\hat{\sigma}_x$ and $\hat{\sigma}_y$, and consider

$$\tilde{x}_j = \frac{x_j - \bar{x}}{\hat{\sigma}_x}, \quad \tilde{y}_j = \frac{y_j - \bar{y}}{\hat{\sigma}_y}, \quad j = 1, \ldots, J. \tag{3}$$

Our strategy is to use the standardized data in (3) in place of the ideal standardized samples in (2) in a two-sample test.

Figure 3 provides examples of how shape differences in X and Y affect the relationship between $\tilde{X}$ and $\tilde{Y}$. The top-left plot shows two distributions that differ in location and scale, but not in shape. For this example, five-hundred samples from each of these distributions serve as data. These data were normalized with (3); the resulting empirical cumulative distribution functions (ECDFs) are displayed in the top-center plot. As can be seen, these ECDFs are quite similar. The bottom-left panel of Figure 3 shows two distributions that differ in location, scale, and shape. Once again, 500 samples from each served as data. The ECDFs of the normalized data are shown in the bottom-center plot. These ECDFs are not as similar as the ECDFs in the top-center panel. When the shapes are the same in data, the normalized distributions are similar; when the shapes are different, the normalized distributions are dissimilar. Therefore, to test for shape differences, we test for differences among the normalized data.

Step # 2: Quantify differences. One plausible statistic for describing the difference between two sets of normalized samples is a sum-squared difference statistic,

$$d^2 = \sum_{j=1}^{J} (\tilde{x}_{(j)} - \tilde{y}_{(j)})^2, \tag{4}$$

where $\tilde{x}_{(j)}$ and $\tilde{y}_{(j)}$ denote the jth order statistic of the respective normalized sample. If we view $\tilde{x}_{(j)}$ and $\tilde{y}_{(j)}$ as sample quantile estimates for the π_jth quantile with $\pi_j = j/(J+1)$, then $\tilde{x}_{(j)} - \tilde{y}_{(j)}$ is the difference in sample quantiles for the two (normalized) samples. The right panels plot these differences as a function of π_j. If distributions X and Y have the same shape, then these segments should be small (top-right panel); as X and Y vary in shape, these segments increase (bottom-right panel). The distance, d^2, is the sum of the squares of these segments.

As an aside, the d^2 statistic is related to the sample correlation between the order statistics of $\mathbf{x}$ and $\mathbf{y}$. In our case, it is easy to show that $d^2 = 2(J-1)(1-r_{xy})$, where r_{xy} is the sample correlation between order statistics. This development is similar to that

underlying the Shapiro-Wilk (1965) one-sample goodness-of-fit test statistic for normality. The Shapiro-Wilk test statistic, denoted r, is the correlation between the order statistics of the sample and the expected value of the order statistics from a standard normal distribution. The shape distance statistic is therefore related to a two-sample expansion of the Shapiro-Wilk test.

The next stage is estimating a sampling distribution for d^2. To our knowledge, this distribution has not appeared in the literature. Thus we implement the test using the bootstrap (Efron, 1979; see Efron & Tibshirani, 1993) to estimate critical values.

Step # 3: Bootstrapping A Sampling Distribution. The bootstrap method is based on resampling from the original data set in some fashion to obtain a so-called "bootstrap" sample. The test statistic is computed for the bootstrap sample. The procedure is repeated many times, and the empirical distribution of the test statistics computed for the bootstrap samples is used to estimate the true sampling distribution of the test statistic.

Let $\mathbf{z}$ be the vector of all the normalized data; i.e., $\mathbf{z} = (\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$. Under the null hypothesis, the elements of $\mathbf{z}$ are approximately independent and identically distributed. (They would be exactly independent if $\mathbf{z}$ were computed with (2) instead of (3).) The simple bootstrap is based on taking M independent bootstrap samples from $\mathbf{z}$ as follows. To obtain the mth bootstrap sample, sample two vectors of data of length J from $\mathbf{z}$ *with replacement*. Suppose these vectors are denoted $\mathbf{x^{(m)}}$ and $\mathbf{y^{(m)}}$, respectively. Regardless of whether the null hypothesis is true or not, the bootstrap samples $\mathbf{x^{(m)}}$ and $\mathbf{y^{(m)}}$ come from the same distribution, hence they have the same shape. Now equations (3) and (4) are applied to each bootstrap sample to compute a bootstrap test statistic, which we will denote d_m^2. Note that d_m^2 is a test statistic computed from two samples from distributions with the same shape, so the null hypothesis is true. The bootstrap idea is to use these samples to estimate the distribution of d^2 under the null.

Step # 4: Estimating the p-value. The collection of d_m^2, $m = 1, \ldots, M$, serves to estimate the sampling distribution of d^2 under shape invariance. If d^2 is in the upper tail of the sampling distribution, the null may be rejected. The bootstrap estimate of the p-value is w/M, where w is the number of bootstrap samples for which $d_m^2 > d^2$. Hence, the decision rule for a desired Type I error rate of α may be constructed by rejecting the null when $p < \alpha$.

An Example

Table 1 provides an example of the bootstrap shape test. A single participant provides 10 observations in each of two conditions. The first two rows, labeled $\mathbf{x}$ and $\mathbf{y}$, display the data. The rows labeled $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ are normalized in accordance with (3). The distance statistic d^2 is computed in accordance with (4); for these data, $d^2 = 1.66$.

The row labeled $\mathbf{x^{(m)}}$ and $\mathbf{y^{(m)}}$ show a bootstrap step; these values were sampled from the concatenation of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ with replacement. Note how values from $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ appear in both $\mathbf{x^{(m)}}$ and $\mathbf{y^{(m)}}$. The following rows show normalized bootstrap samples $\tilde{\mathbf{x}}^{(\mathbf{m})}$ and $\tilde{\mathbf{y}}^{(\mathbf{m})}$. The d^2 statistic for these two vectors is $d_m^2 = 1.49$. We repeated the process for another 999 bootstrap cycles samples and calculated a squared distance for each. The squared distance from the original sample, $d^2 = 1.79$, is less than 634 of the squared distances from the 1000 bootstrap cycles; hence the p-value for the test is .634. For any reasonable α, the null hypothesis of shape invariance cannot be rejected.

Extension of the Shape Test

It is quite straightforward to extend the test to multiple participants. Consider the case in which each participant provides a set of response times in each condition. Let x_{ij} and y_{ij} denote the jth response for the ith participant, $i = 1, ..., I$, in the control and treatment conditions, respectively. We assume that each participant has his or her own

characteristic shift, scale, and shape in each condition. The hypothesis under consideration is that while this shift and scale may vary across conditions and people, shape only varies across people and does not vary across conditions. The first step in the bootstrap test for this hypothesis is to sum square distance over participants:

$$d^2 = \sum_i d_i^2,$$

where d_i^2 is the squared distance statistic for each participant. The same summing is applied to each cycle of the bootstrap; i.e., $d_m^2 = \sum_i d_{i,m}^2$, where $d_{i,m}^2$ is the bootstrapped squared distance for the ith person on the mth cycle.

A second extension covers the case in which there are different numbers of observations in the control and treatment conditions. Two modifications are needed. First, the number of samples comprising $\mathbf{x^{(m)}}$ and $\mathbf{y^{(m)}}$ are adjusted to reflect the numbers of samples in each condition. Second, the distance statistic is modified to account for unequal numbers of samples. Let J^* be the smaller sample size and let $\pi_j = j/(J^* + 1), j = 1, \ldots, J^*$ be an estimate of the CDF of the jth ordered observation for the smaller sample. Next, compute the sequence of π_jth quantiles for $\tilde{\mathbf{x}}^{(\mathbf{m})}$ and $\tilde{\mathbf{y}}^{(\mathbf{m})}$. These sequences are ordered and the test proceeds as before.

Characteristics of the Bootstrap Shape Test

We ran a set of simulations to assess the level (Type I error-rate) and the power of the bootstrap shape test. The reported simulations consisted of 40 participants each observing 200 trials in each of two conditions. These values denote a moderately large-scale experiment in cognitive and perceptual psychology. We repeated this experiment 5000 times with $M = 1000$ resampling cycles.For analyzing data, we recommend a much larger value of M, but the current value provides sufficient accuracy to ascertain the approximate level and power of the tests.

The tests were performed across three distributions that have been useful in modeling RT: the Weibull, the ex-Gaussian, and the inverse-Gaussian or Wald. The density of the ex-Gaussian may be given as

$$f(t;\psi,\theta,\beta) = \theta^{-1}\frac{\exp(z\beta^{-1}) + .5\beta^{-2}}{\beta}\Phi(z-\beta^{-1}), \quad \theta,\beta > 0,$$

where $z = \frac{t-\psi}{\theta}$ and Φ is the cumulative distribution function of the standard normal. Parameters ψ, θ and β serve as location, scale and shape parameters, respectively.

The first set of simulations was performed to assess the level or Type I error-rate of the shape bootstrap test. In this case, true shape values were held constant across people and conditions. These shape values were chosen such that the skewness of the distributions was approximately 0.9, which is representative of observed RT distributions. The Weibull with this skewness is the middle density in Figure 4. In our initial simulations, shifts and scales were varied across each participant-by-item pairing, but this variation had, as expected, no effect on the obtained level and power estimates.

The results of the simulations are shown in Table 2. We report the proportion of times the null was rejected at $\alpha = (.01, .05, .10)$. The first three rows of Table 2 show the case for shape invariance. For these distributions the bootstrap test is conservative for the Weibull and inverse-Gaussian and well-calibrated for the ex-Gaussian.

We ran a second set of simulations to assess power. To model a large shape change, we chose shape parameter values that correspond to a skewness of .7 and 1.1 in the two conditions, respectively. Weibull densities with these skewness are drawn as the outer lines in Figure 4. As shown in the Table, the power is quite high, even when the level is conservative.

Although the bootstrap shape test is useful for the explored RT distributions, one problematic issue is that the level of the test depends on the distribution of the data. We ran an additional set of simulations with the normal, logistic, exponential and Pareto

distributions. The pareto density is given as

$$f(t|\psi,\beta) = \beta\frac{\psi^{\beta}}{t^{\beta+1}}, \quad t \geq \psi, \quad \beta > 0, .$$

where β serves as a shape parameter. As can be seen, the level varies dramatically over these forms. For thin tailed distributions, such as the normal and Weibull with a shape of 1.7, the bootstrap test is highly conservative. It is much better calibrated for exponential tailed distributions (logistic, ex-Gaussian, inverse-Gaussian, exponential) and ridiculously liberal for fat-tailed distributions (Pareto with a shape of 3.0). Fortunately, RT distribution tails tend to be no fatter than an exponential as indicated by either flat or rising hazard functions (Burbeck & Luce, 1982; Wolfe, 1998). Therefore, while the bootstrap shape test is appropriate for RT distributions, it may not be appropriate for other domains.

An Application

To demonstrate the utility of the bootstrap shape test in realistic contexts, we tested empirical data from a priming experiment (Pratte & Rouder, submitted). Participants observed single-digit numbers (2, 3, 4, 6, 7, 8) as targets and had to classify each as being either greater-than or less-than 5. Prior to display of the target, single-digit primes were displayed for 25 ms and masked. These primes, while barely visible, affected the time it took to classify target. Targets preceded by congruent primes, that is, those primes which elicit the same response as the target, were speeded by 13 ms relative to targets preceded by incongruent primes. In Pratte and Rouder's experiment, 43 participants observed 288 congruent and 288 incongruent trials. Pratte and Rouder excluded responses if (a) they were incorrect (5%), (b) response time was outside a 200 ms to 3 s range (.3%), or the prime and target were the same digit (so as to preclude repetition priming effects). After exclusions there were an average of 270 incongruent trials and 184 congruent trials per participant.

Data in priming experiments are often displayed as *delta plots* (e.g., de Jong, Liang, & Lauber, 1994; Ridderinkhof, Scheres, Ooserlaan, & Sergeant, 2005). Delta plots are rotated quantile-quantile (QQ) plots; the diagonal of the QQ plot is rotated to the x-axis (Zhang & Kornblum, 1997). To draw a delta plot, RT deciles are computed for each participant in each condition. For each person and each decile, the difference in RT and the average RT between the incongruent and congruent condition are calculated. The difference score is plotted on the y-axis; the average score is plotted on the x-axis. The left panel of Figure 5 is the delta plot for this experiment. The 43 gray lines in the center plot show this relationship between difference and average RT. The result is that the priming effect is greatest for the quickest responses and falls off as responses slow. The points show averages at each decile. The y-axis value is the average difference across condition; the x-axis value is the average of the average across conditions. The decline in priming effect with slower responses is clear; moreover, this decline has been previously observed in near-liminal number priming experiments (e.g., Greenwald, Abrams, Naccache, & Dehaene, 2003). It is not clear whether the priming effect reverses for the slowest responses. One prevailing interpretation of the decline is that the priming effect is relatively short lived (e.g., Greenwald, Draine, & Abrams, 1996; Neely, 1977). We provide an alternative explanation below after exploring the possibility of shape effects.

The result of the bootstrap shape test is a significant difference in shape across priming conditions ($p < .001$). The right panel of Figure 5 provides a supplemental graphical view of the shape test; the ECDF of all 43 participant's p-values is plotted. Under the shape invariance null, this ECDF should be that of a standard uniform and lie on the diagonal. It does not, indicating values of p_i are less than expected under shape invariance.

The difference in shape across congruent and incongruent conditions is a primary marker of processing. One fruitful avenue to pursue is a mixture explanation. We

speculate that participants are consciously aware of the 25 ms primes for some trials and not for others. Indeed, in a separate block, participants were able to accurately classify the prime as greater-than or less-than 5 on 61% of the trials. Those trials that generate awareness may have a priming effect; e.g., congruent primes that generate awareness result in quicker responses than incongruent primes that generate awareness. Primes that do not generate awareness do not generate a congruency effect. This mixture model would produce shape differences. It is also reasonable to speculate that those primes generating awareness happen on trials in which the participants are paying peak attention to the set of events. Such peak attention will lead to relatively fast responses to the target. This explanation accounts for both the decline in delta plots (congruent effects happen when there is peak awareness and short response times) and a shape change.

Our explanation is admittedly post-hoc and speculative. The exercise demonstrates how shape testing provides important leads for theoretical development. A shape difference is an indicator to the researcher that there is an important marker of processing that needs further exploration. In this case, we would propose a mixture, but clearly more experimentation is needed. A priming researcher may try to localize the components of the mixture though some manipulation, perhaps through manipulating stimulus-to-target asynchronies or the relative proportion of incongruent to congruent primes.

Conclusion

Understanding the shape of RT distributions is critical for linking models of information processing to data. In this paper, we show how parametric tests may not be robust to misspecification and provide a nonparametric alternative. The developed test makes use of modern resampling techniques and is relatively easy to implement. The test proves powerful while maintaining adequate control of Type I error for RT distributions with tails no fatter than an exponential.

There are two large drawbacks to the current test: First, the calibration of the test depends too greatly on the underlying distribution. It would be desirable to have a test whose level was not so dramatically dependent on form of the tail. Second the test applies only to paradigms in which all participants observe stimuli in two conditions. It is not yet evident how to expand the test to factorial designs or continuously measured independent variables. Nonetheless, many experiments in cognition and perception obey the two-condition restriction; in these cases, the bootstrap shape test is appropriate for RT distributions.

References

Anderson, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *The Annals of Mathematical Statistics*, *33*, 255–265.

Ashby, F. G., Tien, J.-Y., & Balikrishnan, J. D. (1993). Response time distributions in memory scanning. *Journal of Mathematical Psychology*, *37*, 526-555.

Burbeck, S. L., & Luce, R. D. (1982). Evidence form auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, *32*, 117-133.

Conover, W. J. (1971). *Practical nonparametric statistics*. New York: Wiley.

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

Dzhafarov, E. N. (1992). The structure of simple reaction time to step-function signals. *Journal of Mathematical Psychology*, *36*, 235-268.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1–26.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.

Greenwald, A. G., Abrams, R. L., Naccache, L., & Dehaene, S. (2003). Long-term semantic memory versus contextual memory in unconscious number processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 235-247.

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science*, *273(5282)*, 1699-1702.

Hockley, W. E. (1984). Analysis of reaction time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *10*, 598-615.

Hsu, Y. F. (1999). *Two studies on simple reaction times: I. On the psychophysics of the generalized Pieron's law. II. On estimating minimum detection times using the time estimation paradigm. Unpublished doctoral dissertation. University of California, Irvine.*

Logan, G. D. (1992). Shapes of reaction time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 883-914.

Lombard, F. (2005). Nonparametric confidence bands for a quantile comparison function. *Technometrics*, *47*, 364-369.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*, 226-254.

Pratte, M. S., & Rouder, J. N. (submitted). *A task-difficulty artifact in subliminal priming.*

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for decisions between two choices. *Psychological Science*, *9*, 347-356.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model analysis of letter masking. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 127-140.

Reed, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer-Verlag.

Ridderinkhof, K. R., Scheres, A., Oosterlaan, J., & Sergeant, J. A. (2005). Delta plots in the study of individual differences: New tools reveal response inhibition deficits in AD/HD that are eliminated by methylphenidate treatment. *Journal of Abnormal Psychology*, *114*, 197-215.

Rouder, J. N. (2000). Assessing the roles of change discrimination and luminance

integration: Evidence for a hybrid race model of perceptual decision making in luminance discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 359-378.

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

Rouder, J. N., Ratcliff, R., & McKoon, G. (2000). A neural network model of priming in object recognition. *Psychological Science*, *11*, 13-19.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 587-604.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 461-479.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics*, *51*, 599-606.

Theeuwes, J. (1994). Stimulus-driven capture and attentional set: Selective search for color and visual abrupt onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 799-806.

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

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response

time models applied to perceptual matching. *Psychonomic Bulletin and Review*, *7*, 208-256.

Wilcox, R. R. (2006). Some results on comparing the quantiles of dependent groups. *Communications in Statistics: Simulation & Computation*, *35*, 893-900.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, *9*, 60-62.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search. *Psychological Science*, *9*, 33-39.

Zhang, J., & Kornblum, S. (1997). Distributional analyses and De Jong, Liang and Lauber's (1994) dual-process model of the Simon Effect. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 1543-1551.

Author Note

Address correspondence to Jeffrey N. Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211 or to rouderj@missouri.edu. This research is supported by NSF grant SES-0351523, NIMH grant R01-MH071418, and ONR grant N00014-06-0106.

Table 1

Sample data, transformed data, and bootstrapped samples provide an example of the shape test.

				$\mathbf{x}$					
0.073	0.58	0.41	0.419	1.107	0.515	0.074	0.233	0.205	0.079
				$\mathbf{y}$					
1.272	0.903	0.524	0.827	0.827	1.308	0.835	0.598	1.087	0.634
				$\tilde{\mathbf{x}}$					
-0.928	0.659	0.127	0.155	2.308	0.455	-0.925	-0.427	-0.515	-0.909
				$\tilde{\mathbf{y}}$					
1.448	0.08	-1.326	-0.202	-0.202	1.581	-0.172	-1.051	0.762	-0.918
				$\mathbf{x}^{(\mathbf{m})}$					
0.762	-1.051	-0.928	0.08	-0.202	0.155	0.455	-0.909	-0.202	0.127
				$\mathbf{y}^{(\mathbf{m})}$					
-0.925	-0.515	0.127	-0.515	0.127	-1.326	0.659	1.581	2.308	-0.202
				$\tilde{\mathbf{x}}^{(\mathbf{m})}$					
1.515	-1.428	-1.229	0.408	-0.05	0.53	1.017	-1.198	-0.05	0.484
				$\tilde{\mathbf{y}}^{(\mathbf{m})}$					
-0.943	-0.577	-0.004	-0.577	-0.004	-1.301	0.47	1.293	1.942	-0.298

Table 2

Observed proportion of rejections of the null hypothesis in simulations.

	Shape Parameter Value		Type I Error			Power		
Distribution	Cond. 1	Cond. 2.	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
			RT Distributions					
Weibull	1.66	1.66	.0012	.0098	.030	-	-	-
Ex-Gaussian	1.19	1.19	.008	.051	.109	-	-	-
inverse-Gaussian	11	11	.003	.022	.061	-	-	-
Weibull	1.48	1.9	-	-	-	.957	.992	.998
Ex-Gaussian	1.0	1.43	-	-	-	.440	.725	.836
Inverse-Gaussian	7.4	18.3	-	-	-	.722	.904	.954
			Other Distributions					
Normal	-	-	.0008	.0044	.015	-	-	-
Logistic	-	-	.005	.040	.098	-	-	-
Exponential	-	-	.015	.080	.159	-	-	-
Pareto	3	3	.72	.86	.92	-	-	-

Figure Captions

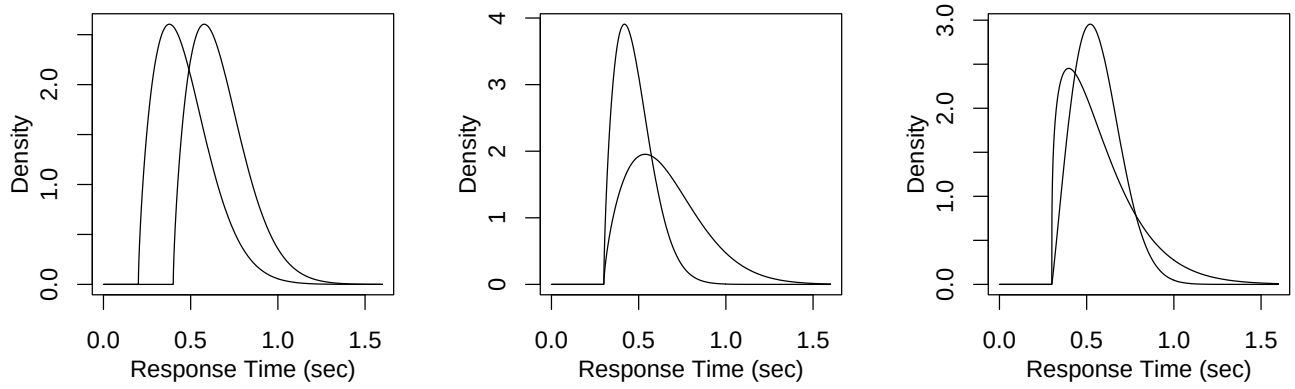
Figure 1. The location-scale-shape taxonomy of differences between distributions. The left, center, and right panels show distributions that differ in location, scale, and shape, respectively.

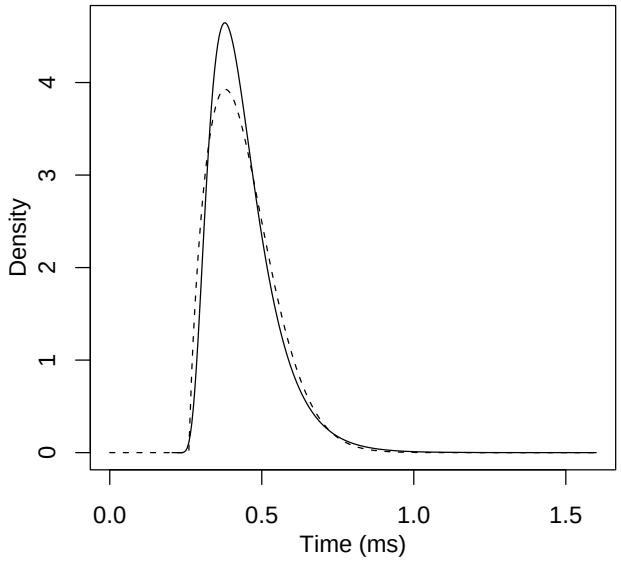
Figure 2. Effects of misspecification on a Weibull parametric test Left: Weibull (dashed) and inverse-Gaussian (solid) pdf's are fairly similar. Right: Cumulative distribution functions of the sampling distribution of log likelihood ratio statistic (G^2) of Weibull test of shape invariance. Lines labeled "W" and "IG" are for Weibull-generated and inverse-Gaussian-generated data, respectively. The dashed line is the theoretical chi-square distribution with 49 *df*. The vertical line indicates the nominal $\alpha = .05$ criterion; the observed Type I error rates are .083 and .853 for the Weibull-generated and inverse-Gaussian-generated data, respectively.

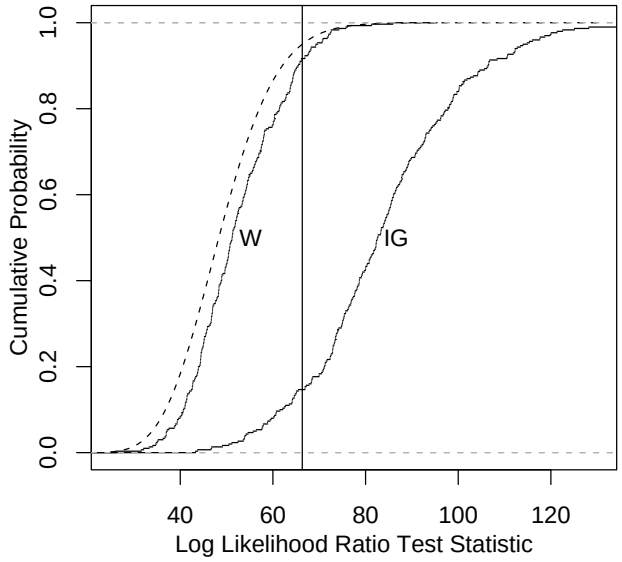
Figure 3. The shape test. Left: Densities underlying data for shape invariance (top row) and shape change (bottom row). Center: Empirical cumulative distribution functions of 500 normalized samples for shape invariance (top row) and shape change (bottom row). Right: Differences in normalized data for shape invariance (top row) and shape change (bottom row).

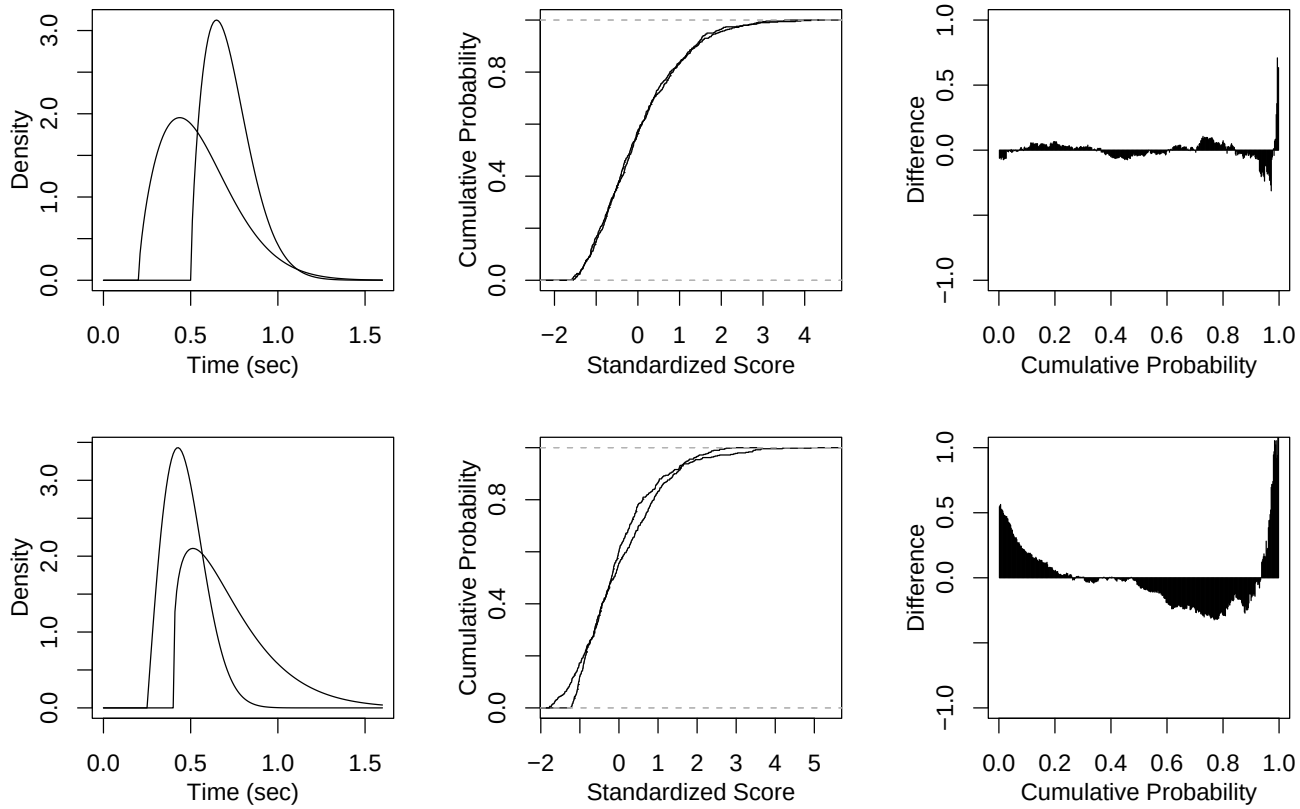
Figure 4. Weibull distributions with skewnesses of .7, .9, and 1.1 (equated means and variances)

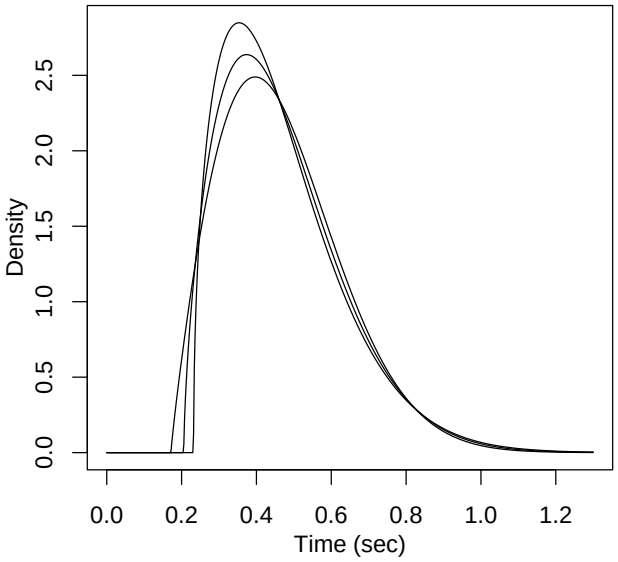
Figure 5. Left: Light gray lines show individual delta plots; solid line is a group average. The late dip below zero is not reliable at the group level. Right: ECDF of p_i. Under the null, the p_i are distributed as a standard uniform (diagonal line). The departure is substantial; shape invariance is rejected.


Density
0.0
1.0
2.0
0.0
0.5
1.0
1.5
Response Time (sec)
Density
0
1
2
3
4
0.0
0.5
1.0
1.5
Response Time (sec)
Density
0.0
1.0
2.0
3.0
0.0
0.5
1.0
1.5
Response Time (sec)

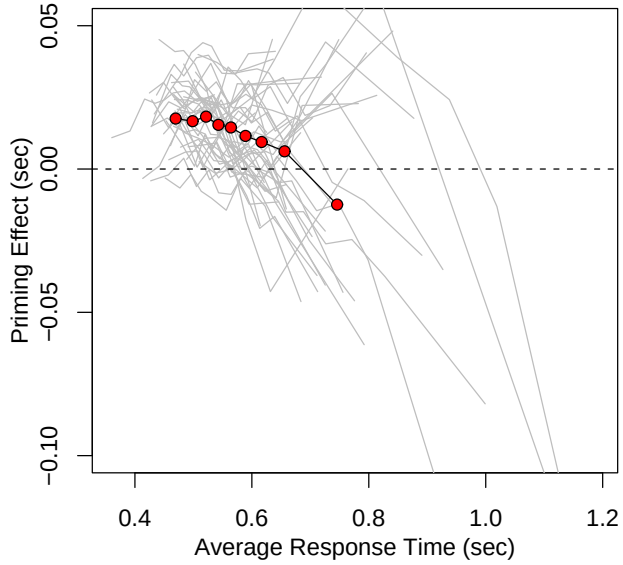

Density
0
1
2
3
4
0.0
0.5
1.0
1.5
Time (ms)

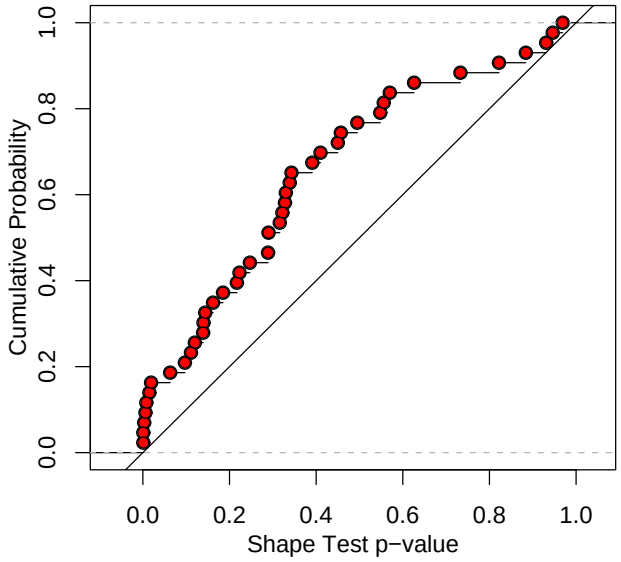

Cumulative Probability
0.0
0.2
0.4
0.6
0.8
1.0
W
IG
40
60
80
100
120
Log Likelihood Ratio Test Statistic


Density
Time (sec)
Cumulative Probability
Standardized Score
Difference
Cumulative Probability
Density
Time (sec)
Cumulative Probability
Standardized Score
Difference
Cumulative Probability


Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Time (sec)

.


Priming Effect (sec)
0.05
0.00
−0.05
−0.10
0.4
0.6
0.8
1.0
1.2
Average Response Time (sec)


Cumulative Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Shape Test p−value