Behavior Research Methods (2021) 53:781–791
https://doi.org/10.3758/s13428-020-01466-5

Percentile rank pooling: A simple nonparametric method for comparing group reaction time distributions with few trials

Jeff Miller[1]

Published online: 31 August 2020

Abstract

Although many studies of reaction time (RT) focus on a single measure of central tendency such as the mean RT, a more detailed picture of the underlying processes can be gained by looking at full distributions of RTs. Unfortunately, for practical reasons it is sometimes difficult to obtain enough trials per participant in a condition of interest to construct such a distribution with existing methods. The purpose of this article is to propose a method of forming group RT distributions that can be used to compare the full distributions of RTs even in an infrequent condition with only a few trials per participant. In brief, the percentile ranks of each participant's infrequent-condition RTs are scored relative to a larger pool including that participant's RTs in other conditions, and a histogram of the infrequent-condition's percentile ranks is then formed by pooling across participants. The resulting histogram of infrequent-condition RT ranks shows where the RTs in that condition tend to fall relative to the other conditions, and this histogram can reveal systematic patterns in the infrequent-condition's RT distribution. To illustrate the method, I present histograms of the ranks of infrequent error RTs (~ 5% of trials), ranked relative to correct responses, in real data sets from Simon and lexical decision tasks.

Keywords Group reaction time distributions · Infrequent conditions

Psychologists have increasingly turned to the study of the distributional properties of the reaction times (RTs) observed in cognitive tasks. A major reason for this trend is that distributional properties can provide useful information beyond what is available in mean RTs (e.g., Luce, 1986). For example, between-condition comparisons of the estimated parameters for specific RT distribution models (e.g., the ex-Gaussian) provide a more nuanced description of experimental effects than a simple comparison of mean RTs (e.g., Balota & Yap, 2011; Heathcote, Popiel, & Mewhort, 1991). Comparisons of RT distributions have also been used extensively to study the time course of experimental effects (e.g., De Jong, Liang, & Lauber, 1994; Reingold, Reichle, Glaholt, & Sheridan, 2012) and to test distributional predictions of RT models (e.g., Miller, 1982; Ratcliff & McKoon, 2008; Ruthruff, 1996).

Because distributional comparisons allow a more in-depth examination of results, several techniques have been developed for the analysis of RT distributions. All of these start by estimating the simple or cumulative distributions of RTs in each single condition (e.g., Van Zandt, 2000), after which it is possible to compare conditions using delta plots (e.g., De Jong et al., 1994), quantile-quantile plots (e.g., Myerson, Adams, Hale, & Jenkins, 2003), estimated divergence points (e.g., Reingold & Sheridan, 2014), or other distribution-based techniques (e.g., quantile regression).

Unfortunately, to study the RT distribution in a given condition, it is usually necessary to have many trials from each participant in that condition (Van Zandt, 2000). For example, hundreds of trials per condition are needed to evaluate full distributional predictions of RT models (e.g., Thomas & Ross, 1980). The ubiquity of individual differences implies that RTs from different participants come from different distributions, however, so RTs cannot be simply pooled together into a "group" distribution as if they were homogeneous (e.g., Engmann & Cousineau, 2011).

Instead of simple pooling, two specialized methods have commonly been used to form a single group RT distribution from the RTs of different participants, and both of these methods estimate the cumulative form of the group distribution. One is quantile averaging (e.g., Ratcliff, 1979; Vincent, 1912). With this method, for each condition, the RTs at a given set of percentiles (e.g., 5%, 15%,...95%) are computed

✉ Jeff Miller
miller@psy.otago.ac.nz

[1] Department of Psychology, University of Otago, PO Box 56, Dunedin 9054, New Zealand

separately for each participant and then arithmetically averaged across participants to obtain the quantile estimate for the overall group distribution[1]. The other method, which might be called "bin averaging" and is commonly used in the analysis of delta plots (e.g., De Jong et al., 1994), is to divide each participant's RTs into bins (e.g., the fastest 10%, the second-fastest 10%, etc.), compute each participant's average RT in each bin, and then average the resulting bin averages across participants. Both of these methods seem to work well when there are many trials per participant in each condition, because having a large number of trials allows the researcher to estimate many percentiles or bin averages for each participant.

The existing methods of forming group distributions are not appropriate, however, when there are relatively few trials per condition. With both quantile and bin averaging, the number of trials per participant limits the resolution of the estimated distribution. With ten trials per participant, for example, the researcher can estimate at most ten percentiles (i.e., 5, 15,...95) or ten bin averages per participant, and of course the resolution is even poorer with fewer trials than that. The problem of small trial numbers is further aggravated if the number of trials in the condition varies across participants. When that happens, different participants provide estimates of RTs at different percentile ranks or bins, so there are no common points at which to compute averages. For example, a participant with ten RTs provides estimates at the percentile ranks of 5, 15, 25,...95, whereas a participant with nine RTs provides estimates at the ranks of 6, 17, 28, 39,...94. Since these percentile ranks do not match, it is not clear how to get percentile averages. This is not an uncommon problem, because the number of trials in a condition might vary across participants either because occasional trials are lost or excluded (e.g., outlier RTs, equipment malfunction) or because the condition is at least partly defined by factors beyond the researcher's control (e.g., erroneous responses).

The purpose of this article is to suggest a very simple and flexible nonparametric method of "percentile rank pooling" that can be used to examine group RT distributions for conditions in which there are only a few trials per participant[2]. The method aims to fill the gap where existing methods are difficult or impossible to use, providing researchers with a new tool to investigate RT distributions in conditions that were previously unexamined for lack of adequate numbers of trials. As will be seen, percentile rank pooling can reveal interesting patterns in participants' underlying true RT distributions, even when there are only a few RTs per participant and even when there are substantial individual differences between participants in the location, scale, skew, and even the shape of RT distributions.

Using percentile ranks to normalize across participants

In percentile rank pooling, each RT from each participant is normalized by computing its percentile rank, RT_{pr}, within a large comparison set of RTs from that participant. Histograms of the ranks are then tabulated by simply pooling the ranks across all participants, separately for each condition. The resulting histograms show where the RTs in each condition tend to fall relative to the other conditions. As is illustrated later with examples, the appropriate set of comparison RTs for computing percentile ranks will depend on the research question, and it could include RTs from all of the participant's trials or just the trials from a relevant subset of conditions.

Within a given set of comparison trials, the percentile rank of each RT value t is computed as

$$RT_{pr}(t) = \frac{L + 0.5 \times E}{N} \times 100 \qquad (1)$$

where L is the number of RTs in the comparison set that are less than t, E is the number of RTs equal to t, and N is the total number of RTs in the comparison set. If there are 100 RTs in the comparison set and no ties, for example, then the percentile ranks of the RTs, from smallest to largest, will be 0.5, 1.5, 2.5,...99.5. A few of these percentile ranks would reflect RTs from trials in the infrequent condition of interest, whereas the rest would reflect the other RTs in the comparison set.

The rationale for converting to percentile ranks is that these are necessarily in a comparable 0–100 range for all participants, even if the original RTs of the different participants were quite different due to individual differences in mean, standard deviation, skewness, or even distribution shape (e.g., ex-Gaussian, gamma, lognormal). It thus is reasonable to pool percentile ranks across participants, even though it would be inappropriate to pool their actual RTs. In this sense, transforming to percentile ranks works much like transforming scores on different scales to Z-scores for comparison, but without the implicit assumption of a symmetric, approximately normal distribution.

By definition, the histogram of percentile ranks across the whole comparison set is necessarily uniform (i.e., all percentile ranks are equally frequent) for each individual participant and thus also for the ranks pooled across participants. For

[1] Cousineau, Thivierge, Harding, and Lacouture (2016) suggested a variant of this method in which the RTs of individual participants at each percentile are transformed and geometrically averaged, arguing that this variant may produce group distributions that are more similar to the form of the underlying individual-participant distributions, which are not necessarily reproduced with standard quantile averaging (Thomas & Ross, 1980).

[2] Other nonparametric techniques have also been suggested for examining RT distributions (e.g., Lombardi, D'Alessandro, & Colonius, 2019; Maris & Maris, 2003), but these require many trials per participant and have been focused on testing a specific model prediction known as the race model inequality (Miller, 1982) rather than on obtaining a general picture of RT distributions.

example, suppose there are 100 RTs per participant in the comparison set, with no ties. For each participant, one of the 100 RTs is the fastest and receives the lowest percentile rank (0.5%), another is the second fastest and receives the second-lowest percentile rank (1.5%), and so on up to the slowest RT receiving the largest percentile rank (99.5%). It follows that each of the possible percentile ranks is obtained once for each participant, and the resulting pooled percentile ranks are perfectly uniform across the 0–100 range. The principle is the same even if there are different numbers of trials for the different participants: Each participant will have one RT at each of the percentile ranks possible with their own value of N, so every participant's percentile ranks will be spread evenly across the 0–100 range, and the resulting pooled histogram will also have evenly spread percentile ranks.

Although the overall histogram of *all* of the pooled percentile ranks must necessarily be uniform, the histograms of percentile ranks in *each condition* will generally not be. To the extent that condition A is faster than condition B, for example, condition A will tend to have a preponderance of the smaller pooled percentile ranks and B will have a preponderance of the larger ones. (More complex histogram patterns will be illustrated shortly.) Thus, differences in the distributions of the pooled percentile ranks in the two conditions will tend to reflect systematic differences in their underlying RT distributions. If the two conditions do happen to have identical RT distributions, however, then any RT is equally likely to come from either condition, and both conditions will necessarily have uniform distributions of pooled percentile ranks.

Using the ex-Gaussian distribution as a convenient model of observed RTs (e.g., Hohle, 1965; Luce, 1986), Fig. 1 illustrates how percentile rank pooling can reveal the general features of the RT distribution in the infrequent condition relative to those of the frequent condition and can thereby provide new information about the infrequent condition's RT distribution. Suppose that the overall mean RT in a frequent condition (occurring in 90% of trials) is 100 ms faster than in an infrequent one (10% of trials). At the distributional level, this difference in means might arise in any of the three different ways shown in Fig. 1a–c. In Fig. 1a, the mean RT difference primarily reflects a difference at the fastest end of the RT distributions, with very few or none of the fastest trials coming from the infrequent condition, as would be produced by a 100-ms increase in the μ parameter of the ex-Gaussian in the infrequent condition. For example, fewer than 2% of the trials with RTs less than 500 ms come from the infrequent condition, although 10% of all trials come from that condition. Alternatively, in Fig. 1c, the difference between conditions is mainly due to the stretched upper tail of the infrequent condition distribution (i.e., increased τ parameter), so the increase in mean RT is largely due to an increase in the proportion of very slow trials. For example, more than 20% of the trials with RTs greater than 1000 ms come from the infrequent condition, as compared with 10% of all trials. Figure 1b represents a compromise between these two extremes, with half of the mean RT difference associated with a shift in the infrequent distribution (i.e., increased μ) and half associated with a stretch in its upper tail (i.e., increased τ).

Figure 1d–f shows that percentile rank pooling can accurately recover these true differences between the pairs of underlying distributions shown in Fig. 1a–c. To obtain the displayed histograms of pooled percentile ranks in a concrete and easily visualized way, I simulated 10,000 participants for each pair of underlying RT distributions shown in Fig. 1a–c[3]. Two hundred simulated RTs were generated for each participant, and all RTs were generated independently either from the frequent condition, with probability 90%, or from the infrequent one, with probability 10%. The 200 simulated RTs of each participant were converted to percentile ranks using Eq. 1 with $N = 200$, and these percentile ranks were pooled across all participants to form separate histograms for the frequent and infrequent conditions.

Several patterns are visible in the pooled percentile rank histograms of Fig. 1d–f. First, the total frequencies shown in the histograms are much higher for the frequent condition than for the infrequent one, which was of course inevitable because there were 90% and 10% of trials in these two conditions, respectively. Second, the histograms for the frequent conditions are nearly uniform. As was mentioned earlier, the overall histogram of percentile ranks is necessarily uniform (i.e., pooling across conditions), and the histogram of the frequent condition cannot deviate too much from that because it includes most of the trials. Thus, percentile rank pooling provides little information with respect to the distribution of RTs in the frequent condition, and it would be more informative to examine these distributions with a traditional method such as quantile averaging.

Third and most importantly, the infrequent-condition histograms of Fig. 1d–f clearly differentiate between the different types of distribution-level effects shown in Fig. 1a–c. In Fig. 1d, the virtual absence of infrequent RTs with percentile ranks in the 0–20 range makes it clear that this condition is slower than the frequent one because it produces hardly any of the fastest responses. In contrast, there are many infrequent RTs with low percentile ranks in Fig. 1f, but here the infrequent condition has a disproportionately large number of quite slow

[3] These percentile rank histograms can also be computed directly with a numerical approximation, as is described in the Appendix, and the dotted lines show the approximations obtained in this manner.

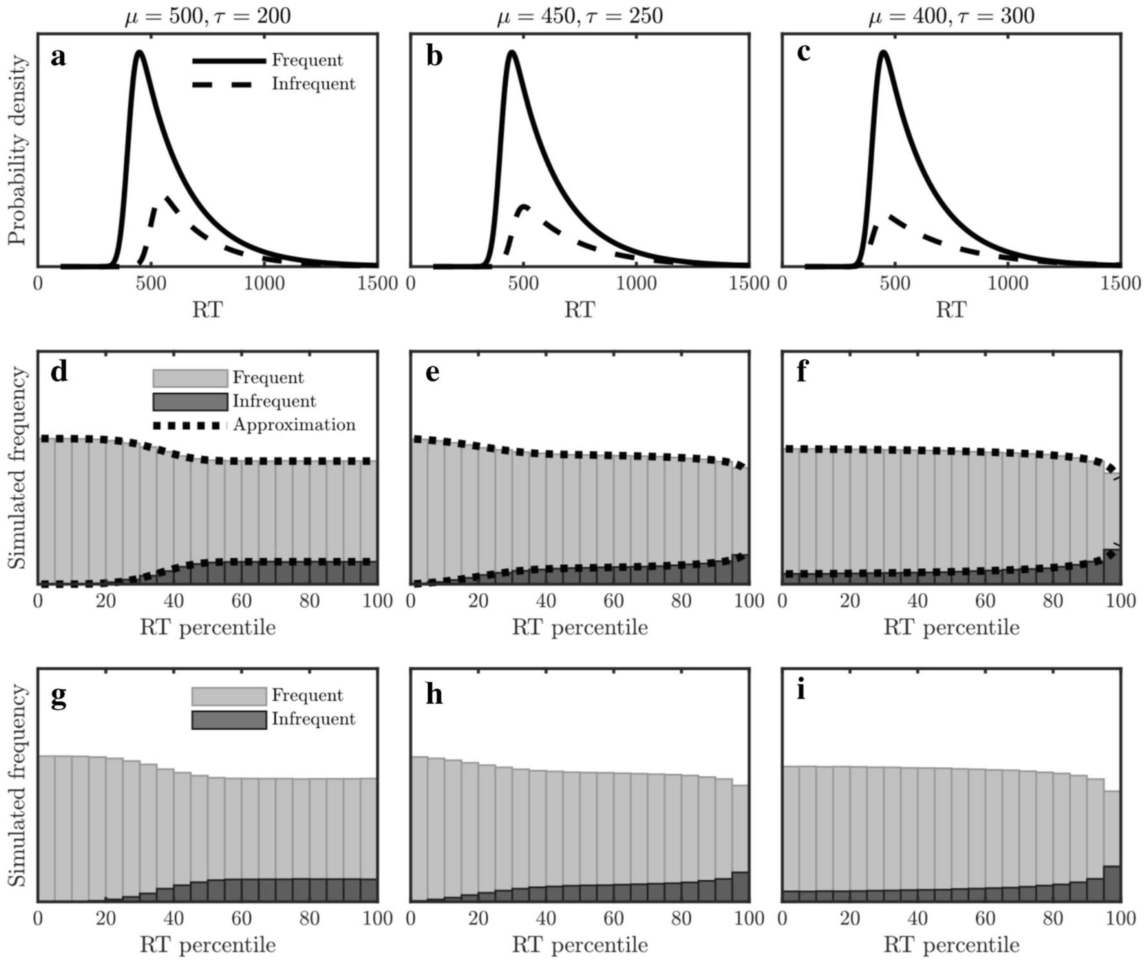

Fig. 1 True underlying distributions of RT (**a**–**c**) and simulated histograms of percentile ranks (**d**–**i**) for frequent (90%) and infrequent (10%) conditions. All of the underlying true RT distributions were ex-Gaussian, and the distribution parameters in the frequent condition were always μ = 400 ms, σ = 30 ms, and τ = 200 ms. The distribution parameters of the infrequent condition in each column are shown above panels **a**–**c**. In **a**–**c**, the densities for the frequent and infrequent conditions have been scaled to 75% and 25% of their true values to reflect the difference in simulated trial frequency. This is not to scale—the actual simulated frequencies were 90% and 10%, but the infrequent distribution is difficult to see when the distributions are shown to that scale. Panels **d**–**f** show histograms of RT percentiles computed from RTs simulated according to the distributions shown in **a**–**c**. The dotted lines in **d**–**f** show approximations of the expected histograms computed as described in the Appendix. Panels **g**–**i** show histograms of RT percentiles computed from RTs simulated with random variation among participants in the parameters of the underlying RT distributions (see text for details)

RTs (i.e., percentile ranks of approximately 80–100)[4]. The infrequent condition of Fig. 1e shows a pattern intermediate between these two extremes, corresponding to the compromise between shift and stretch effects in the infrequent distribution of Fig. 1b.

Figure 1g–i shows that percentile rank pooling can also recover the main features of the infrequent-condition RT distribution when there is participant-to-participant variation in the underlying RT distributions. These histograms were obtained by simulations analogous to those used to produce Fig. 1d–f, except that the parameters of the true underlying frequent and infrequent RT distributions were generated randomly for each participant. Specifically, for the frequent condition, each participant's ex-Gaussian μ parameter was randomly selected from a normal distribution with mean 400 ms and standard deviation 40 ms, and each participant's τ parameter was randomly selected from a normal distribution with mean 200 ms and standard deviation 20 ms. In addition, for each participant the amount of slowing in the infrequent condition was randomly selected from a normal distribution with mean 100 ms and standard deviation 10 ms, and this slowing either increased the infrequent condition's μ parameter (Fig.

[4] The complementary relationship between the frequent and infrequent histograms can also be seen in Fig. 1d–f. Specifically, for each RT percentile, there is a fixed sum of the simulated histograms for the frequent and infrequent conditions, so if the simulated count in the infrequent condition increases at a given RT percentile, the simulated count in the frequent condition must correspondingly decrease. This complementary relationship is simply a consequence of the fact that the overall distribution of RT percentile ranks must be uniform pooling across conditions.

1g) or τ parameter (Fig. 1i), or had half of its effect on each of these (Fig. 1h). Once the parameters of a simulated participant's frequent and infrequent RT distributions had been randomly selected, that participant's RTs for the frequent and infrequent conditions were randomly generated and analyzed just as in Fig. 1d–f. Importantly, despite the participant-to-participant parameter variation, the percentile ranks of the simulated RTs again clearly show different distributional-level slowing effects comparable to those in Fig. 1d–f.

In sum, percentile rank pooling can provide information about the RT distribution in an infrequent condition. By pooling this condition's percentile ranks across many participants, the researcher can build up a fine-grained distribution-level picture of how the RTs in this condition relate to the RTs in a frequent condition. As was explained in the introduction, it would be difficult or impossible to get a comparably fine-grained distributional picture with a standard technique like quantile averaging because of the small numbers of trials per participant in this condition, especially if this number varied across participants.

The nonparametric (i.e., rank-based) character of percentile rank pooling implies that it is robust with respect to variation in the parameters and even the shapes of the underlying RT distributions. This is because the method depends directly only on the theoretical distributions of percentile ranks (i.e., analogous to the histograms shown in Fig. 1d–f), not on the underlying distributions of RTs that give rise to those theoretical distributions (i.e., analogous to those shown in Fig. 1a–c). This means that the pooled histogram of percentile ranks will reflect any pattern that is common across the true percentile rank distributions of the individual participants, even if those participants have radically different RT distributions—potentially even more different than those simulated in Fig. 1g–i. For example, if all participants had a disproportionately large number of infrequent-condition RTs with percentile ranks in the range of 50–70, then the pooled percentile rank histogram would likewise have a high proportion of ranks in this range, regardless of the similarity of the underlying individual-participant RT distributions. This would be true even if different participants had different shapes of underlying RT distributions—for instance, some ex-Gaussian, some ex-Wald, some lognormal, some gamma, etc.—because the technique uses only nonparametric information about percentile ranks.

Illustration: Error RTs in Simon tasks

To evaluate the method's performance with respect to real data, I began with a situation for which some information about the true underlying distribution is independently available—the Simon task (for a review, see e.g., Hommel, 2011). As is explained below, existing evidence has some implications concerning the distribution of error RTs in this task, even though errors are infrequent.

To review briefly, in a common visual version of the Simon task, participants must respond with the left or right hand depending on the color of a stimulus square (e.g., Mittelstädt & Miller, 2020). Each stimulus square is presented to the left or right of fixation, with location being irrelevant to the correct response. Responses are faster and more accurate when the stimulus location is congruent with the required response hand (i.e., square on the left requiring a left-hand response) than when it is spatially incongruent (i.e., square on the left requiring a right-hand response). Moreover, when the difference in accuracy (congruent minus incongruent) is plotted as a function of RT (an accuracy "delta plot"; e.g., De Jong et al., 1994; Dittrich, Kellen, & Stahl, 2014), the accuracy advantage for the congruent condition decreases as RT increases. This pattern implies that errors must tend to be especially fast in the incongruent condition, and it is of interest to see whether percentile rank pooling would also yield that conclusion.

To check the pattern of Simon task error RTs with percentile rank pooling, I analyzed the RT data of Experiments 1–4 of Mittelstädt and Miller (2020). Each of these experiments compared congruent versus incongruent RTs under two conditions differing in task difficulty. For example, in Experiment 1 the color discrimination was easy or difficult, and in Experiment 4 the responses were made with the hands (easy) or feet (difficult). The type of difficulty manipulation had little effect on the distribution of error RTs, so the results shown here are collapsed across the four experiments. For each participant in each experiment, percentile ranks of all RTs were obtained separately for the four combinations of congruent/incongruent and easy/difficult conditions, which had overall error rates of 3.9% (congruent, easy), 5.9% (congruent, difficult), 6.1% (incongruent, easy), and 7.5% (incongruent, difficult). Histograms of the resulting percentile ranks for correct and error trials are shown in Fig. 2.

The histograms for the correct trials are nearly uniform, as was necessarily the case since these were most of the trials entered into the percentile ranking. The histograms for errors are much more interesting. In the incongruent conditions, the histograms of error RT percentile ranks appear somewhat bimodal, with a large mode of fast RTs (percentile ranks ≈ 0–20) and a smaller mode of slow RTs (percentile ranks ≈ 90–100). The large mode of fast RTs is consistent with the previously reported pattern of a larger congruency effect on the accuracy of faster RTs mentioned earlier, so the method successfully recovered this expected pattern, which would also be evident in a comparison of correct versus error mean RTs. The increased error rate for the slowest incongruent responses has not previously been discussed, however. Interestingly, the congruent conditions also show a slightly increased number of errors in the highest RT percentile ranks, just like the incongruent ones. The preponderance of slow errors in both

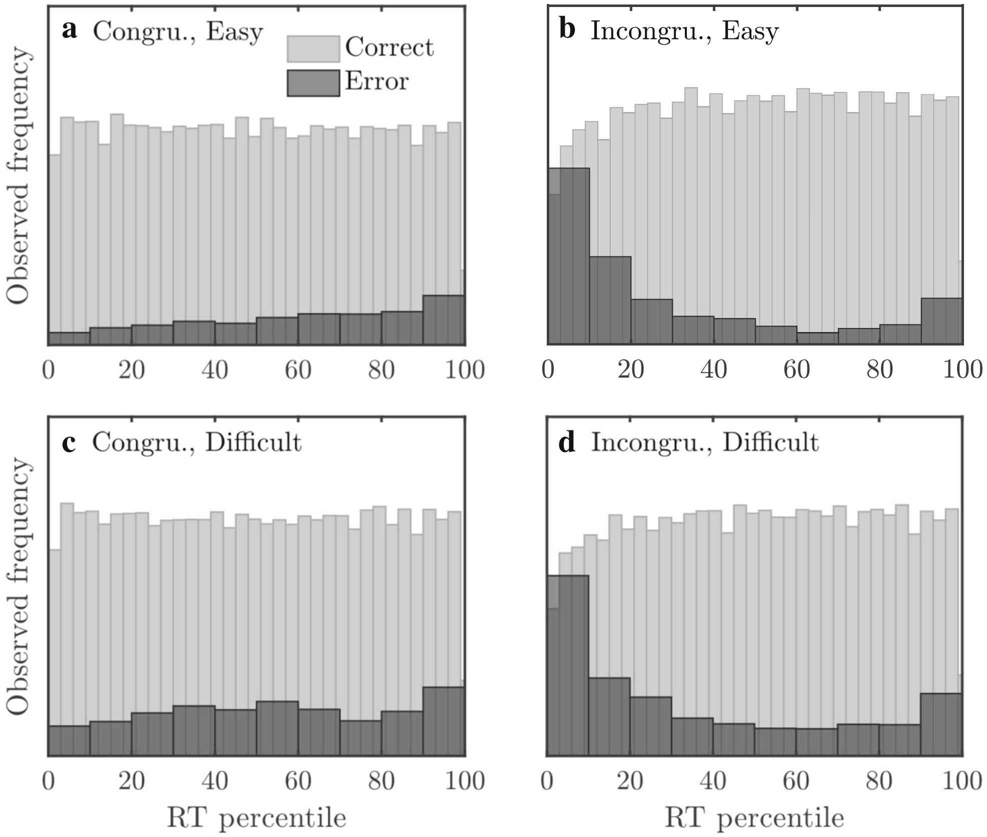

Fig. 2 Frequency distributions of RT percentile ranks of correct and error trials in the congruent and incongruent conditions of the Simon tasks reported in experiments 1–4 of Mittelstädt and Miller (2020). The experiments used different manipulations of task difficulty, but the results have been pooled across experiments because these produced no clear differences in the histograms of RT percentiles. **a** Congruent trials, easy task. **b** Incongruent trials, easy task. **c** Congruent trials, difficult task. **d** Incongruent trials, difficult task

congruent and incongruent conditions would not be evident in the usual accuracy delta plots if the increased error rates for slow responses were about equal between these conditions, because these plots are only sensitive to the accuracy difference between conditions. Thus, the percentile rank pooling method not only reveals the pattern seen in delta plots but also provides additional information. In this case, the mode for slow errors suggests that some condition-independent source of trial-to-trial variability tended to make responses both quite slow and also inaccurate.

Illustration: Error RTs in lexical decision tasks

As another illustration of the percentile rank pooling method with real data, I applied it to the RT distributions of errors in lexical decision tasks using the large, publicly available data sets of Ferrand et al. (2010) and Hutchison et al. (2013). After excluding some problematic participants and RTs (e.g., RTs less than 200 ms), these data sets included 1,862,117 and 828,963 trials, respectively, from 944 and 503 participants, with error rates of 7.61% and 4.06%. Responses were substantially faster to word stimuli than nonwords in both data sets, so I computed RT percentile ranks separately for these two stimulus types for each participant. The histograms of RT percentile ranks for trials with correct responses versus errors were then tabulated across participants, and the results are shown in Fig. 3.

Again, the histograms for the correct trials are nearly uniform, as they had to be. The histograms of error RTs are not uniform, and in fact they appear to differ for words versus nonwords. For words, errors were most often relatively slow, suggesting that participants sometimes considered words for a relatively long time before incorrectly concluding that they were nonwords.

For nonwords, in contrast, the histograms of RT percentile ranks for errors were again bimodal, like those seen in the incongruent trials of the Simon task. The mode of fast errors to nonwords could be due to the presence of some highly word-like nonwords in the stimulus sets, and these nonwords could presumably be identified by examining the error rates and average RT percentile ranks for the individual nonword stimuli. Alternatively, the fast errors to nonwords could reflect fast guesses of the "word" response, in which case they would be distributed randomly across the different nonwords. In that case, the absence of comparable fast errors to word stimuli

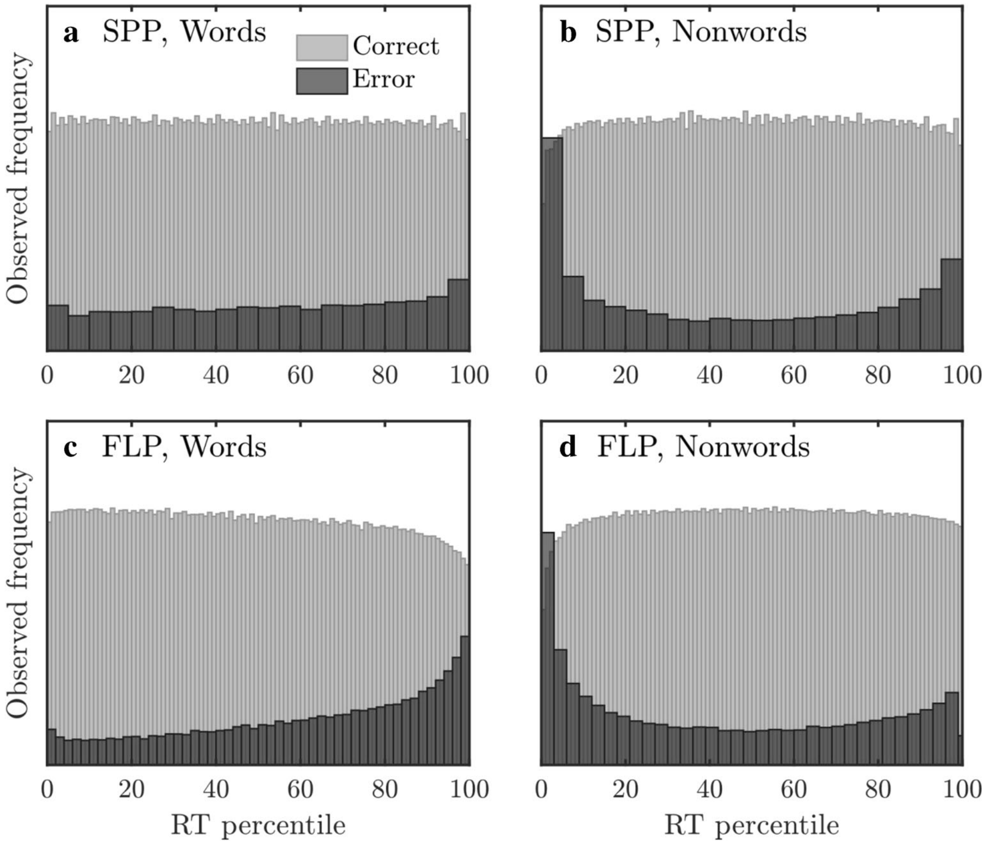

Fig. 3 Frequency distributions of RT percentile ranks of correct and error trials in lexical decision tasks for trials with word and nonword stimuli in the data sets from the Semantic Priming Project (SPP; Hutchison et al., 2013) (**a**, **b**) and the French Lexicon Project (FLP; Ferrand et al., 2010) (**c**, **d**)

would indicate that there were many fewer fast guesses of the "nonword" response.

There also appears to have been a small mode of slow errors for the nonwords, at least relative to the frequency of errors with percentile ranks in the range of approximately 30–70%. Like the slow errors to word stimuli, these could reflect stimulus items about which participants had considerable doubt. Again, it would be illuminating to examine the specific nonword stimuli with relatively high error rates and relatively high RT percentile ranks.

Converting RT percentile ranks back to RTs

Although histograms of percentile ranks convey a great deal of distributional information, for some purposes it might be useful to estimate the RT distributions of the frequent and infrequent conditions in terms of actual RTs rather than in terms of the RT percentile scores RT_{pr} produced by the percentile rank pooling method. For example, one might want to test an RT model predicting the forms of the correct and error RT distributions.

The percentile rank pooling method can be extended to generate RT scores in the original millisecond scale. This requires the additional step of estimating a group-average cumulative RT distribution function, $F(t)$, which requires some parametric assumptions. Once that distribution has been determined—by whatever means—the RT percentile ranks of the frequent and infrequent conditions can be converted back onto the millisecond scale via the inverse function F^{-1}.

To illustrate this process, I used the RTs to nonword stimuli from the Semantic Priming Project that were used to construct Fig. 3b. To estimate the group-average RT distribution parametrically, I first estimated the ex-Gaussian parameters μ, σ, and τ for all nonword RTs (i.e., corrects and errors combined) of each participant separately. I then estimated the overall group distribution as the ex-Gaussian with the means of the estimated individual-participant's μ, σ, and τ parameters. The regular and cumulative probability densities of this estimated group distribution are shown as the solid lines in Fig. 4a and b. I checked this estimated group distribution by computing the corresponding quantile-averaged cumulative distributions, shown as the dotted line in Fig. 4b. Visually, the ex-Gaussian and quantile-averaged group distributions are in very good agreement, suggesting that either would be a reasonable estimate for the group distribution. I used the ex-Gaussian distribution estimate because it was more convenient computationally.

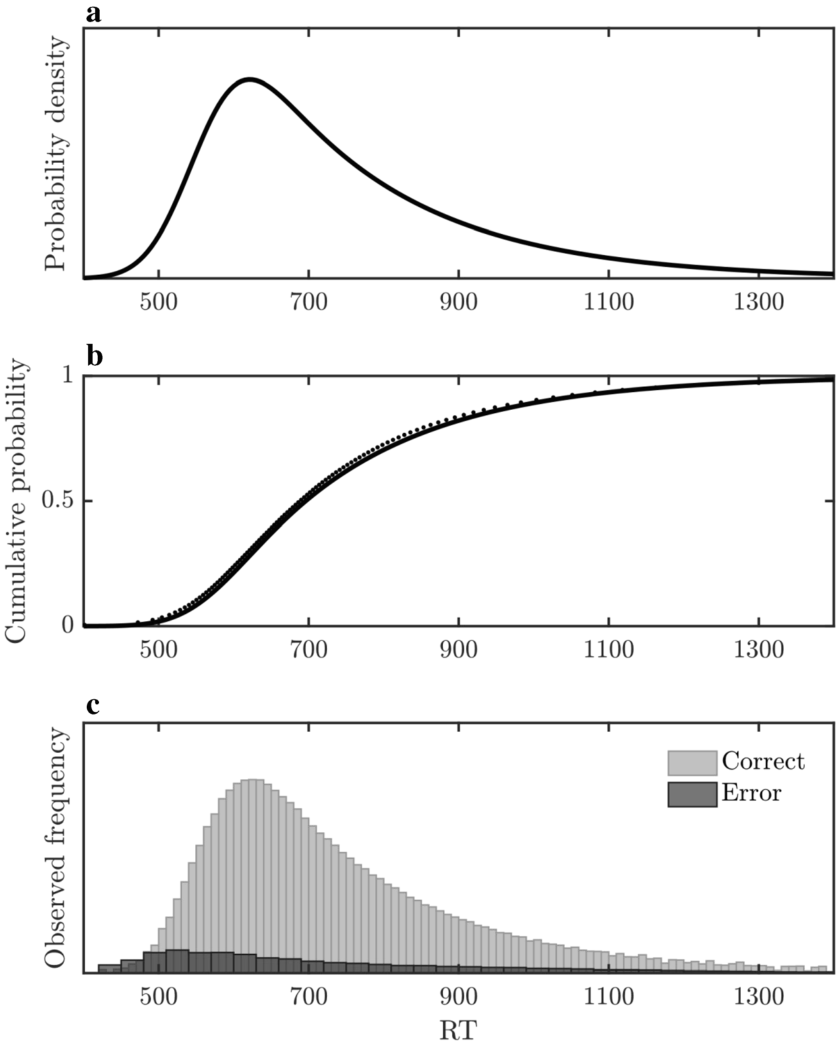

Fig. 4 Illustration of converting RT percentile ranks into RTs using an estimated group-average RT distribution. Regular (**a**) and cumulative (**b**) group-average RT distributions computed from the RTs to nonwords in the data from the Semantic Priming Project (SPP; Hutchison et al., 2013). The solid lines represent the group distribution formed by fitting ex-Gaussian distributions to each participant's RTs individually and averaging the fitted parameters across individuals. It is an ex-Gaussian distribution with μ = 552.65, σ = 51.48, and τ = 197.06. The dotted line represents the cumulative group distribution formed by quantile averaging. **c** Frequency distributions of correct and error RTs derived from the percentile ranks shown in Fig. 3b by retrieving the group-average RT score with the corresponding percentile

Using the group-average cumulative ex-Gaussian distribution $F(t)$ just estimated, I computed the underlying RT associated with each of the percentile ranks RT_{pr} shown in Fig. 3b using the inverse cumulative distribution function $F^{-1}(RT_{\mathrm{pr}})$. These RTs were then tabulated separately for correct and error trials, and the resulting histograms are shown in Fig. 4c. These histograms could be used, for example, to test RT models predicting specific shapes of the correct and error RT distributions. Note also that the preponderance of fast errors seen in the RT percentile ranks of Fig. 3b is quite visible in the RT values of Fig. 4c, but the preponderance of slow errors is not. Evidently, the spread in the long tail of the RT distribution makes it harder to compare the conditions with respect to their slow RTs in the millisecond scale. This difficulty is not present with the RT percentile ranks because these are bounded at 100%, effectively eliminating the stretch in the upper tail. Thus, percentile rank pooling may be especially helpful in checking for between-condition differences in particularly slow responses.

General discussion

The percentile rank pooling method involves computing the percentile ranks for each individual's RTs pooled together across a relevant set of to-be-compared conditions. These percentile ranks are then pooled across participants and tabulated into histograms separately for each condition. The resulting histograms show where the ranks of the RTs in the different conditions tend to fall relative to one another. These histograms can reveal systematic differences between the RT percentile ranks of conditions at any point(s) in the distributions of ranks, including conditions with only a few trials per participant.

Percentile rank pooling is proposed as a supplement to—not a replacement for—existing methods of examining RT distributions. When researchers have enough RTs to estimate the distribution for each participant in each condition, comparisons of these distributions in their original millisecond units would seem to allow simpler interpretations than percentile-ranked distributions. As mentioned earlier, simple or cumulative distributions can be compared directly (e.g., Miller, 1982; Ruthruff, 1996), or these can serve as the basis for comparing delta plots (e.g., De Jong et al., 1994), quantile-quantile plots (e.g., Myerson et al., 2003), or estimated divergence points (e.g., Reingold et al., 2012; Reingold & Sheridan, 2014, 2018). When there are not enough RTs to accurately estimate the distribution in each condition, however, percentile rank pooling seems promising as an alternative method for gaining insight into the differences between distributions.

In this article, I have focused on comparing the RT histograms of frequent and infrequent conditions, because the new method seems to be the first allowing examination of RT distributions in conditions with so few RTs per participant. The method was illustrated by examining the RT histograms of errors in several data sets, because errors are infrequent in most RT experiments, yet their RT distributions can be theoretically informative. As noted earlier, distributions of error RTs cannot be computed by standard methods like quantile averaging, because these methods require more trials per participant.

Percentile rank pooling could also potentially be useful in a number of other situations where it is difficult to get large numbers of trials per participant in a condition of interest. For example, in some studies the nature of the research question dictates that relatively small numbers of trials will be available in a particular condition. In studies manipulating stimulus probability, response probability, or expectancy, for example, it is inherent in the experimental manipulation that there are relatively few trials per participant in low-probability and unexpected conditions (e.g., Crossman, 1953; Hyman, 1953; Katzner & Miller, 2012; Klein, 1994; Mattler, 2003; Miller & Pachella, 1973; Starns & Ma, 2018). Furthermore, in addition to errors, certain interesting types of trials may be uncommon even when they are obtained in high-probability conditions (e.g., trials at low levels of practice). In conditions with relatively few trials, RT distributions have rarely if ever been examined with existing methods, and percentile rank pooling could provide new information.

The percentile rank pooling method might also be useful for comparing RT histograms in data sets with many infrequent conditions and no frequent one, which can arise in studies with multi-factor within-subject designs. For example, with a 2 × 3 × 3 × 4 design, it would usually be difficult to get more than approximately 10 trials per participant in each condition, which would at best provide only a crude picture of the within-condition RT distributions. Computing each RT's percentile rank pooled across all conditions, however, provides a percentile-based normalization of each participant's RTs relative to their full set of RTs for the experiment as a whole. Pooling these percentile ranks across participants to form condition-specific percentile rank histograms could provide a detailed picture of where the RTs in each condition fell in comparison to all of the other conditions.

One clear limitation of the percentile rank pooling method is that it can only be used for within-subject comparisons. The RTs of the conditions being compared must be ranked after combining all of the scores together, and this is only appropriate if they come from the same participant.

A second limitation is that the method is only useful when the RT distributions of different conditions overlap. If all of the RTs in one condition are faster than all of the RTs in the other, then the pooled percentile ranks for one condition would be uniform within the range of 0–P%, and those of the other condition would be uniform in the range of P–100%, where P is the proportion of trials in the faster condition[5]. In such cases, pooling would not provide new information about the shapes of the RT distributions in the two conditions. Given the large natural trial-to-trial variability of RTs for a given participant, however, observed RT distributions in different conditions virtually always overlap, so the percentile rank pooling method will nearly always provide some further information about distribution shapes.

A third limitation of the method is that there is currently no simple approach for testing the statistical significance of any patterns that might appear in the pooled percentile rank distributions. Each participant provides multiple RT_{pr} values to each constructed histogram (e.g., Fig. 2), so the values are not all independent of one another, making it problematic to use standard techniques for distributional comparisons (e.g., Engmann & Cousineau, 2011). Bootstrapping and randomization techniques appear to be promising approaches for significance testing, though these would have to be adapted on a case-by-case basis. For the error trials shown in Fig. 3b, for example, 95% bootstrap confidence intervals[6] show no overlap in bin proportions among any of the top three bins or any of the bottom three bins, suggesting that the preponderance of quite fast and quite slow errors is a real phenomenon rather than a statistical fluke. There may also be situations in which inferential techniques are unnecessary, of course, such as when there are strong patterns

[5] The boundary at P would be step-like if the proportion of trials in each condition was fixed experimentally (e.g., stimulus probability), but there would be some smearing at the boundary if the proportion in each condition varied across participants (e.g., correct versus errors).

[6] Computed by sampling participants with replacement.

in a large data set or when the same patterns are replicated across multiple experiments. Even without statistical testing, the method can also be used for descriptive purposes and exploratory research. Furthermore, certain types of descriptive distributional results may point toward more focused follow-up analyses for which standard inferential techniques can be used even with small numbers of trials (e.g., comparisons of RT variability).

Acknowledgements This article was written while the author was on sabbatical leave at the Institut für Psychologie (Allgemeine Psychologie) of the University of Freiburg, and I thank Prof. Andrea Kiesel for the facilities provided there. I also thank Ludovic Ferrand for providing the raw data from the French Lexicon Project. I am grateful to Patricia Haden, Victor Mittelstädt, Wolf Schwarz, and two anonymous reviewers for helpful comments on earlier versions of the article.

Open practices statement The data sets analyzed for the Simon and lexical decision tasks were retrieved from the online repositories indicated in the original publications of Ferrand et al. (2010), Hutchison et al. (2013), and Mittelstädt and Miller (2020).

Compliance with ethical standards

Conflict of interests The author declares that he has no conflicts of interest with respect to the authorship or publication of this article.

Appendix

An Approximation to the Distribution of Percentile Ranks As defined in the main text (Eq. 1), the distribution of pooled percentile ranks is a discrete distribution with N possible values, where N is the number of RTs pooled together for computation of percentile ranks. This appendix shows how to compute a continuous approximation to this discrete distribution. The approximation is a convenient alternative to tabulating percentile rank distributions by simulation for any specific set of assumptions about the underlying RT distributions.

Assume that $k \geq 2$ conditions are pooled together for the computation of percentile ranks, with p_i being the proportion of trials in condition i, $i = 1...k$. Assume further that the RTs in condition i come from a probability distribution with cumulative probability function F_i. The overall distribution of the pooled RTs is thus a mixture distribution F_M, with

$$F_M(t) = \sum_{i=1}^{k} p_i \cdot F_i(t) \tag{2}$$

(e.g., Everitt & Hand, 1981)[7].

We seek the probability distribution of the percentile ranks of the RTs in condition i, the cumulative form of which can be denoted as $F_{pr,i}(r)$.

$$\begin{aligned} F_{pr,i}(r) &= \Pr\left(RT_{pr,i} \leq r\right) \\ &= \Pr\left(F_M(RT_i) \leq r\right) \\ &= \Pr\left(RT_i \leq F_M^{-1}(r)\right) \\ &= F_i\left(F_M^{-1}(r)\right) \end{aligned} \tag{3}$$

Thus, for any given set of conditions occurring with probability p_i and having RT distributions F_i, a continuous approximation to each condition's distribution of pooled percentile ranks can be computed numerically.

References

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*(3), 160–166. https://doi.org/10.1177/0963721411408885

Cousineau, D., Thivierge, J.-P., Harding, B., & Lacouture, Y. (2016). Constructing a group distribution from individual distributions. *Canadian Journal of Experimental Psychology*, *70*(3), 253–277. https://doi.org/10.1037/cep0000069

Crossman, E. R. F. W. (1953). Entropy and choice time: The effect of frequency unbalance on choice-response. *Quarterly Journal of Experimental Psychology*, *5*, 41–51. https://doi.org/10.1080/17470215308416625

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 & Performance*, *20*, 731–750. https://doi.org/10.1037/0096-1523.20.4.731

Dittrich, K., Kellen, D., & Stahl, C. (2014). Analyzing distributional properties of interference effects across modalities: Chances and challenges. *Psychological Research*, *78*(3), 387–399. https://doi.org/10.1007/s00426-014-0551-y

Engmann, S., & Cousineau, D. (2011). Comparing distributions: The two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test. *Journal of Applied Quantitative Methods*, *6*(3), 1–17.

Everitt, B. S., & Hand, B. J. (1981). *Finite mixture distributions.* London, England: Chapman & Hall.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., … Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496. https://doi.org/10.3758/BRM.42.2.488

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response-time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*, 340–347. https://doi.org/10.1037/0033-2909.109.2.340

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*, 382–386. https://doi.org/10.1037/h0021740

Hommel, B. (2011). The Simon effect as tool and heuristic. *Acta Psychologica*, *136*(2), 189–202. https://doi.org/10.1016/j.actpsy.2010.04.011

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. (2013). The Semantic

[7] The term "mixture distribution" is usually used to describe models in which each trial's condition is determined randomly (e.g., correct versus error), so that the number of trials in each condition varies randomly. This random variation is not important for the present purposes, however, because Eq. 2 also describes the overall RT distribution if the number of trials in each condition is fixed by the researcher (e.g., high versus low probability).

Priming Project. *Behavior Research Methods*, *45*(4), 1099–1114. https://doi.org/10.3758/s13428-012-0304-z

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*, 188–196. https://doi.org/10.1037/h0056940

Katzner, S., & Miller, J. O. (2012). Response-level probability effects on reaction time: Now you see them, now you don't. *Quarterly Journal of Experimental Psychology*, *65*(5), 865–886. https://doi.org/10.1080/17470218.2011.629731

Klein, R. M. (1994). Perceptual-motor expectancies interact with covert visual orienting under conditions of endogenous but not exogenous control. *Canadian Journal of Experimental Psychology*, *48*, 167–181. https://doi.org/10.1037/1196-1961.48.2.167

Lombardi, L., D'Alessandro, M., & Colonius, H. (2019). A new nonparametric test for the race model inequality. *Behavior Research Methods*, *51*(5), 2290–2301. https://doi.org/10.3758/s13428-018-1170-0

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* Oxford, England: Oxford University Press.

Maris, G., & Maris, E. (2003). Testing the race model inequality: A nonparametric approach. *Journal of Mathematical Psychology*, *47*, 507–514. https://doi.org/10.1016/S0022-2496(03)00062-2

Mattler, U. (2003). Combined perceptual or motor-related expectancies modulated by type of cue. *Perception & Psychophysics*, *65*, 649–666. https://doi.org/10.3758/BF03194589

Miller, J. O. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*(2), 247–279. https://doi.org/10.1016/0010-0285(82)90010-X

Miller, J. O., & Pachella, R. G. (1973). Locus of the stimulus probability effect. *Journal of Experimental Psychology*, *101*(2), 227–231. https://doi.org/10.1037/h0035214

Mittelstädt, V., & Miller, J. (2020). Beyond mean reaction times: Combining distributional analyses with processing stage manipulations in the Simon task. *Cognitive Psychology*, *119*(101275). https://doi.org/10.1016/j.cogpsych.2020.101275

Myerson, J., Adams, D. R., Hale, S., & Jenkins, L. (2003). Analysis of group differences in processing speed: Brinley plots, Q-Q plots, and other conspiracies. *Psychonomic Bulletin & Review*, *10*(1), 224–237. https://doi.org/10.3758/BF03196489

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*, 446–461. https://doi.org/10.1037/0033-2909.86.3.446

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Reingold, E. M., Reichle, E. D., Glaholt, M. G., & Sheridan, H. (2012). Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. *Cognitive Psychology*, *65*(2), 177–206. https://doi.org/10.1016/j.cogpsych.2012.03.001

Reingold, E. M., & Sheridan, H. (2014). Estimating the divergence point: a novel distributional analysis procedure for determining the onset of the influence of experimental variables. *Frontiers in Psychology*, *5*, 1432. https://doi.org/10.3389/fpsyg.2014.01432

Reingold, E. M., & Sheridan, H. (2018). On using distributional analysis techniques for determining the onset of the influence of experimental variables. *Quarterly Journal of Experimental Psychology*, *71*(1), 260–271. https://doi.org/10.1080/17470218.2017.1310262

Ruthruff, E. D. (1996). A test of the deadline model for speed-accuracy tradeoffs. *Perception & Psychophysics*, *58*(1), 56–64. https://doi.org/10.3758/BF03205475

Starns, J. J., & Ma, Q. (2018). Response biases in simple decision making: Faster decision making, faster response execution, or both? *Psychonomic Bulletin & Review, 25*(4), 1535–1541. https://doi.org/10.3758/s13423-017-1358-9

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. https://doi.org/10.1016/0022-2496(80)90003-6

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review*, *7*, 424–465. https://doi.org/10.3758/bf03214357

Vincent, S. B. (1912). The function of the viborissae in the behavior of the white rat. *Behavioral Monographs*, *1* (No. 5).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.