Refining the Law of Practice

Nathan J. Evans[a], Scott D. Brown[b], Douglas J. K. Mewhort[c] and Andrew Heathcote[db]

[a] Department of Psychology, Vanderbilt University, USA

[b] School of Psychology, University of Newcastle, Australia

[c] Department of Psychology, Queen's University, Ontario, Canada

[d] Division of Psychology, University of Tasmania, Australia

Correspondence concerning this article may be addressed to: Nathan Evans, Department of Psychology, Vanderbilt University, Nashville TN, USA; Email: nathan.j.evans@uon.edu.au

Abstract

The "Law of Practice" – a simple nonlinear function describing the relationship between mean response time (RT) and practice – has provided a practically and theoretically useful way of quantifying the speed-up that characterizes skill acquisition. Early work favored a power law, but this was shown to be an artifact of biases caused by averaging over participants who are individually better described by an exponential law. However, both power and exponential functions make the strong assumption that the speedup always proceeds at a steadily decreasing rate, even though there are sometimes clear exceptions. We propose a new law that can both accommodate an initial delay resulting in a slower-faster-slower rate of learning, with either power or exponential forms as limiting cases, and which can account for not only mean RT but also the effect of practice on the entire distribution of RT. We evaluate this proposal with data from a broad array of tasks using hierarchical Bayesian modeling, which pools data across participants while minimizing averaging artifacts, and using inference procedures that take into account differences in flexibility among laws. In a clear majority of paradigms our results supported a delayed exponential law.

Keywords:

Law of practice — Learning — Skill Acquisition — Bayesian hierarchical models

Introduction

Across a broad array of tasks, practice leads to a speed-up that is monotonic, but non-linear, with early large gains later giving way to diminishing returns. Whether the function relating mean response time (RT) to practice has a single simple mathematical form has been the subject of much debate, focusing mainly on power vs. exponential functions. Both functions have a long history of support (Crossman, 1959; Newell & Rosenbloom, 1981; Seibel, 1963; Snoddy, 1926; Suppes, Fletcher, & Zanotti, 1976; Heathcote, Brown, & Mewhort, 2000; Josephs, Silvera, & Giesler, 1996; Rosenbloom & Newell, 1987), and continue to influence theorizing and analysis in many fields, such as motor ability (Yarrow, Brown, & Krakauer, 2009), memory (Pavlik & Anderson, 2005; Roediger, 2008a), complex task learning (Lee & Anderson, 2001), and sub-task learning (Anglim & Wynton, 2015), just to name a few.

As typically applied to practice data, power (Equation 1) and exponential (Equation 2) functions describe the change in mean RT as a function of practice, N (usually measured by number of trials), in terms of three estimated parameters that quantify: 1) mean asymptotic performance ($\alpha\prime$), 2) the amount by which initial performance is slower than the asymptote (β), and 3) the rate of learning (r):

$$\alpha\prime + \beta N^{-r} \tag{1}$$

$$\alpha\prime + \beta e^{-rN} \tag{2}$$

Figure 1 illustrates the two functions and their fit to two sets of simulated data: one from the power function, and one from the exponential function. Both functions potentially provide a useful "measurement model", compactly describing the salient features of practice data in terms of a small number of reliably estimated parameters. The parameters $\alpha\prime$ and β are practically useful, quantifying, respectively, the best possible level of performance after extended practice and the gain in performance that can be achieved by practice (relative to trial zero for the exponential function and trial one for the power function). The r parameter relates to the theoretically crucial difference between these functions in terms of their "relative learning rates", the current rate of improvement divided by the amount of potential remaining improvement. For the exponential function, the relative learning rate is a constant, r, so each extra practice trial improves performance by a *constant proportion* of the amount of possible improvement that remains. For the power function, the relative learning rate is r/N, so the amount learned from each trial is a *decreasing proportion* of the remaining available improvement, indicating a depletion or exhaustion in the ability to learn, such as when there are initially multiple learning processes operating at different time scales, with fast scale processes completing early in practice so only the slow scale processes are responsible for improvements later in practice (Newell & Rosenbloom, 1981).

Early empirical investigations of the law of practice usually averaged data over participants, and consistently favored a power function (e.g., Crossman, 1959; Seibel, 1963; Snoddy, 1926; Suppes et al., 1976). In a seminal paper, Newell and Rosenbloom (1981) examined 15 experiments using a wide variety of tasks including motor-skills,

elementary perceptual and memory decisions, complex routines, and problem solving. After averaging each data set across participants, they found that the power function provided a better fit than the exponential function for all 15 data sets. This strong regularity was very influential, with the power function taking on the status of a behavioral law. This "Power Law of Practice" guided many subsequent cognitive theories, including Instance Theory (Logan, 1988, 1992), transfer of automaticity (Kramer, Strayer, & Buckely, 1990), parallel distributed processing models (Cohen, Dunbar, & McClelland, 1990), and cognitive architectures such as SOAR (Laird, Newell, & Rosenbloom, 1987) and ACT-R (J. R. Anderson & Lebiere, 1998).

In the decades that followed statistical approaches to model selection advanced considerably, leading Heathcote et al. (2000) to re-assess the evidence with improved methods. Most importantly, they fit practice functions to data from each subject and condition (where a condition was sometime a single item) separately, because averaging over exponential functions can approximate a power function when – as is invariably the case – there are individual differences in learning rates (R. B. Anderson & Tweney, 1997; Brown & Heathcote, 2003; Estes, 1956; Myung, Kim, & Pitt, 2000), or differences between conditions. For example, in the case of individual differences, a power function can emerge because all learners contribute to improvements in the average function early in practice, whereas only slow learners contribute later in practice. Heathcote et al. found that practice curves were better fit by the exponential function in the majority of participants and conditions across 17 data sets from all areas of experimental psychology, including tasks such as visual and memory search, mental arithmetic, alphabetic arithmetic, counting, and motor

learning. Given the consistency of their results, Heathcote et al. suggested repealing the Power Law of Practice in favor of an Exponential Law of Practice.

Our article has two main aims. First, we extend the previous exponential and power laws in two ways, to account for slower learning early in practice, and to account for the entire distribution of RT. It has been reported (e.g., Rickard, 1997) that there can be an initial delay in learning, sometimes followed by very fast learning (Rickard, 2004). Neither the exponential nor power laws can account for these findings, so we extend them to do so. Our second extension not only provides a description of the variability which is always observed in performance, and its decrease with practice (Logan, 1992), it also links the laws to a process-based explanation of that variability – the Lognormal Race evidence accumulation model (LNR; Heathcote & Love, 2012). Our second aim is to re-assess the evidence for the exponential and power laws, and our new extensions of them. This assessment uses state-of-the-art model-selection methods that take account of differences in flexibility due to their different mathematical forms. This allows the investigation of whether learning is best described by a power or exponential law, as well as how widespread is the occurrence of learning delays.

Transitions in Practice Functions

It has long been recognized (e.g., Bryan & Harter, 1897, 1899) that learning curves can sometimes display one or more plateaus – regions where little improvement occurs – or even periods of deteriorating performance typically associated with

fatigue. Such phenomena are of considerable applied relevance, but it has also long been known that under conditions that both allow for rests, and give initial training instructions that provide the optimal strategy – conditions typical of controlled experimental settings – such plateaus tend to disappear (e.g., Keller, 1958). However, the strong assumption made by power and exponential functions that the rate of learning always decreases over the course of practice fails to account for a phenomenon that can remain even in optimal experimental settings: an initial period of slower learning followed by a speed-up before the final approach to asymptote. **This S-shaped pattern has been attributed to a transition between performance based on an algorithm and direct retrieval, with the transition either mediated by a race between (Logan, 1988), or probabilistic mixture of (Rickard, 1997), these two strategies.**

This failure compromises the usefulness of a "Law of Practice" as a measurement model, because it cannot describe potentially important variations in the transition process. It may also compromise its utility in terms of theoretical implications through introducing systematic biases. For example, a power function speedup is a core theoretical tenant of Instance Theory, derived from a direct retrieval mechanism based on race between memory instances. As a power function has an initial very rapid decrease, it will be disadvantaged relative to alternatives with a slower initial decrease (e.g., an exponential function) in fits to data that contain a flat initial region due to a slow algorithm-to-retrieval transition.

Rather than trying to address this issue with a process model, which typically requires estimation of many additional parameters (and so can compromise desir-

able measurement model properties such as providing a simple and stably estimated summary), we propose the following new practice functions, with a single additional parameter, $\tau \geq 0$:

$$\alpha\prime + \beta\frac{\tau + 1}{\tau + N^r} \tag{3}$$

$$\alpha\prime + \beta\frac{\tau + 1}{\tau + e^{rN}} \tag{4}$$

When $\tau = 0$ the new functions reduce to the standard power and exponential forms. As τ increases an initial period of slow decrease emerges, delaying the transition to power or exponential learning. Hence, we call these new forms the "delayed-power" and "delayed-exponential" functions, and the τ parameter as a "delay" parameter.

When both τ and the learning rate (r) are large the initial slow period is followed by a rapid decrease. This regime has the potential to model the cases where, after a period of slow performance and little improvement, participants gain insight into the availability of a more efficient strategy, whose application results in a rapid transition to improved performance followed by more gradual learning before an asymptote is achieved. For example, Rickard (2004) found rapid transitions of this sort for some items in a difficult alphabet arithmetic problems (e.g., verifying that F + 9 = 0 is true, where the addition signifies the number of letters to be counted in alphabetical order from the "F" to check whether it results in the letter on the right, "O").

Figure 2 illustrates the descriptive flexibility of the delayed functions with fits

of the exponential and delayed-exponential functions to simulated data from four different data-generating processes: the standard exponential (top left), the delayed exponential (bottom left), and on the right two cases mimicking data patterns reported by (Rickard, 2004). On the top right there is an initial period of slow constant performance that then becomes a faster standard exponential (similar to "Type 2" items in Rickard's Figure 6). On the bottom right the data were generated in the same way but with a small probability of samples from the initial slow performance occurring later (similar to "Type 1" items in Rickard's Figure 5). Note that neither data generating process directly corresponds to a fitted function in these latter two cases. The exponential function is so tightly constrained that it struggles to predict any data other than its own, with over-estimation where initial learning is slow. More importantly, it also shows distortions in later regions where it is the true model. In contrast, the delayed exponential is able to provide a good account of almost all of the data. It can, of course, accommodate standard exponential function data, as it is a special case, but also data where learning is initially slow or completely absent. Its only failing is evident in the lower right panel, where it over-estimates the central tendency of the bulk of the data late in practice due to contamination by occasional very slow responses.

Accounting for the RT Distribution

There are a number of strong and general regularities that might be described as "Laws" of RT measurement. First, RTs are variable, so even when the same response is made under nearly identical conditions the observed RT can vary quite widely. RT almost always has a unimodal positively skewed distribution, and that

distribution has a lower bound greater than zero below which responses cannot be stimulus-contingent (Luce, 1986). Further, changes in mean RT are accompanied by proportionate changes in its standard deviation, both in general (Wagenmakers & Brown, 2007), and in the particular domain of practice effects (Logan, 1992). The practice laws described so far are descriptively inadequate as they do not take account of these pervasive regularities. This flaw also means it is unclear how to fit the functions to data, as fitting must take account of variability. For example, the most commonly used method, unweighted least squares (e.g., Heathcote et al., 2000), makes RTs early in practice overly influential, as they should be down-weighted due to their greater variability.

We propose to approximate all of the afore-mentioned regularities while requiring the estimation of only two extra parameters. In particular, we assume a shifted multiplicative lognormal model of the RT distribution, where the practice function is multiplied by a lognormally distributed random variable, and then positively shifted by some constant that reflects the lower bound of response time for stimulus contingent responses. A random variable has lognormal ($\mathcal{LN}$) distribution if its logarithm is normally distributed. The lognormal distribution is unimodal and positively skewed, and has a long history of use in RT analyses, both as a measurement model (see, e.g., Woodworth & Schlosberg, 1954) and more recently as a cognitive-process model (Heathcote & Love, 2012; Rouder, Province, Morey, Gomez, & Heathcote, 2014; Terry et al., 2015; Ulrich & Miller, 1993). We parameterized the lognormal distribution in terms of the mean and standard deviation of the underlying normal distribution; to avoid ambiguity we refer to them, respectively, as

the "log-mean" (μ) and "log-sd" (σ). The log-mean uniquely determines the median of the longormal distribution and log-mean and log-sd together determine its moments (e.g., mean, standard deviation, skew etc.). The only extra parameter to be estimated is σ, as the log-mean is given by the practice functions already described.

We also estimate a shift parameter ($t_0 \geq 0$) corresponding to the fastest possible RT, which we assume does not change with practice. As shown in the following equations (which specify the practice functions as random variables), we partitioned the mean asymptote parameters ($\alpha\prime$) in Equations 3 and 4 into constant (t_0) and variable ($\alpha = \alpha\prime - t_0$) asymptote parameters. The variable parts of the equation (i.e., α plus the practice function) are multiplied by a lognormal random variable with a log-mean fixed at zero (and hence a median fixed at one) and log-sd of σ. As a consequence, median RT equals Equations 3 or 4. We describe these functions as the "delayed random power function" and the "delayed random exponential" function.

$$RT \sim t_0 + \left(\alpha + \beta \frac{\tau + 1}{\tau + N^r}\right) \mathcal{LN}(0, \sigma) \tag{5}$$

$$RT \sim t_0 + \left(\alpha + \beta \frac{\tau + 1}{\tau + e^{rN}}\right) \mathcal{LN}(0, \sigma) \tag{6}$$

Even though σ is fixed across practice, the standard deviation of RT decreases in proportion to the decrease in both median and mean RT with practice (once t_0 is subtracted; Figure 1, bottom row). This occurs because both the mean and standard deviation of the lognormal distribution are proportional to e^{μ}. Hence, the practice

function for mean RT (with t_0 subtracted) and for the standard deviation of RT are exactly proportional to the median practice function (again with t_0 subtracted).

The addition of an explicit model of variability to the practice functions affords a comprehensive and practically useful characterization of behavior. For example, it supports quantification of best-case performance in terms of the lower percentiles of RT distribution, at the start, during, and after extensive practice. Similarly, it supports quantification of worst-case performance throughout practice in terms of the higher percentiles, enabling quantitative management of risks associated with slow responses. This addition also allows easy computation of the likelihood of any set of parameters given the data. The likelihood quantifies all of the information in the data relevant to model estimation. Thus, optimal fitting methods are supported, such as maximum likelihood estimation and the method we adopt, hierarchical Bayesian estimation.

Hierarchical Bayesian Estimation and Model Flexibility

We fit hierarchical models because they enable enhanced measurement by pooling information over participants without introducing the distortions associated with fitting to averaged data. They also avoid the implausible assumption, implicit in individual fitting, that estimates from one participant provide no constraint on estimates for other participants (Shiffrin, Lee, Kim, & Wagenmakers, 2008, see Evans & Brown, 2017b and Evans, Rae, Bushmakin, Rubin, & Brown, 2017, for applications). Bayesian methods make estimation of hierarchical models easier in practice. More importantly, they also allow us to address the important issue of model flexibility.

More flexible models usually provide a better fit to data than simpler mod-

els, regardless of whether they reflect the data-generating process or whether they provide the best predictions of future data. Basic approaches to this issue, such as AIC and BIC, balance goodness-of-fit against model flexibility measured by the number of estimated parameters (e.g., Akaike, 1974; Schwarz, 1978; Burnham & Anderson, 2004). However, these basic approaches fail to take account of differences in complexity due to differences in functional form, which occur because the increase in flexibility endowed by adding a parameter depends on the mathematical form of a parametric model (Myung, 2000; Myung & Pitt, 1997; Evans & Brown, 2017a; Evans, Howard, Heathcote, & Brown, 2017). In the domain of forgetting curves, for example, Averell and Heathcote (2011) found that the addition of an asymptote parameter made exponential functions were more flexible than power functions, whereas the opposite held with no asymptote. Hence, it is possible that the advantage for the exponential over power function found by Heathcote et al. (2000) may be an artifact of greater flexibility. The issue of functional-form complexity is also paramount in determining whether the addition of a delay parameter is justified.

Here, we address model flexibility using the Widely Applicable Information Criterion (WAIC: Vehtari, Gelman, & Gabry, 2017; Gelman, Hwang, & Vehtari, 2014) to select the model that is best in the sense of being able to predict future data. WAIC is more stable to compute than DIC, which has been used in the past for hierarchical models (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). It is an approximation to leave-one-out cross validation, that is, selecting the model that, when fit to all but one data point, and on average over all possible such exclusions, best predicts the left out data.

In the next section we apply these methods to fit and compare the two models specified in Equations 5 and 6. We also compare these delayed laws to the simpler power and exponential laws without delays by fixing $\tau = 0$. We analyzed data from the same 17 experiments as Heathcote et al. (2000). In addition, we also analyzed Rickard's (2004) data in order to test whether the delay functions could account for rapid transitions.

Method

Data

Table 1 provides a brief description of each data set examined by Heathcote et al. (2000); more details are provided by Heathcote et al., and the original sources. These data are available online (`http://www.newcl.org/node/6`).

Data Analysis

In order to minimize the influence of assumptions about the psychological effects of experimental manipulations, we allowed all five parameters for the random power and exponential practice functions, t_0, α, β, r, and σ, as well as the τ parameter for their delayed forms, to differ across experimental conditions in each data set. **Different test items were treated as different conditions within our analyses, to avoid the potential biases that can result from averaging over different learning functions with different rates.** We assumed a hierarchical structure in which each parameter in each experiment – and where applicable each group within each experiment – was drawn from independent normal population dis-

tributions, each with separate mean (m) and a standard deviation (s) parameters. The population distributions were truncated to positive values in accordance with the definition of the τ and σ parameters, and to enforce non-negative RT values and non-increasing effects of practice: $t_0 \sim \mathcal{TN}(m_{t_0}, s_{t_0})$, $\beta \sim \mathcal{TN}(m_\beta, s_\beta)$, $r \sim \mathcal{TN}(m_r, s_r)$, $\sigma \sim \mathcal{TN}(m_\sigma, s_\sigma)$, $\alpha \sim \mathcal{TN}(m_\alpha, s_\alpha)$, and $\tau \sim \mathcal{TN}(m_\tau, s_\tau)$. We chose vague priors for the m and s parameters, having, respectively, positive truncated normal ($\mathcal{TN}$) and gamma (Γ) distributions: $m_{t_0} \sim \mathcal{TN}(1, 1)$, $m_{r,exponential} \sim \mathcal{TN}(0.2, 0.2)$, $m_{r,power} \sim \mathcal{TN}(0.5, 0.5)$, $m_\alpha, m_\beta \sim \mathcal{TN}(1, 2)$, $m_\sigma \sim \mathcal{TN}(2, 2)$, $m_\tau \sim \mathcal{TN}(0, 3)$, and $s_{t_0}, s_r, s_\alpha, s_\beta, s_\sigma, s_\tau \sim \Gamma(1, 1)$

We used the Metropolis-Hastings algorithm with proposals generated by the differential evolution algorithm (Turner, Sederberg, Brown, & Steyvers, 2013) to obtain posterior parameter samples. Differential evolution involves running multiple parallel Markov chains, and generating the new parameter proposal for each chain by adding the existing value to the difference between the values of two other randomly selected chains. The generation of proposals also involves two tuning parameters. One is a small random perturbation added to all proposals, which was set at 10^{-7}. The other, a "step length" measure, was set according to the dimension of the subject-level parameter space, at $\frac{2.38}{\sqrt{2k}}$, where k is the dimension. For each data set we sampled $5k$ Markov chains, and for each chain we generated 19,000 samples as burn-in. Posterior convergence was confirmed by visual inspection. In the vast majority of posteriors (99.9%) the $\hat{R}$ measure of chain mixing (Brooks & Gelman, 1998) was below 1.1 (the generally recommended maximum value), and in all cases it was less than 1.25. After burn-in, we drew 21,000 samples from each chain, and thinned

to use every 10^{th} iteration for further analysis. Data points from all subjects in each data set were combined in order to calculate a single WAIC value for each data sets. We report WAIC on the deviance scale, meaning that, as with AIC/BIC/DIC, lower WAIC values indicate a better model.

We also fit another variant of the power function, termed the "general power function" by Heathcote et al. (2000), which was originally proposed by Newell and Rosenbloom (1981). The general power function includes an additional parameter, P, which allows for learning before the start of the experiment [1]:

$$\alpha\prime + \beta(N + P)^{-r} \tag{7}$$

As discussed by Heathcote et al. (2000), with high values of P this function can have slower rate of learning early in practice than the regular power function, and mimic the regular exponential function to some extent. However, this function is unable to produce the S-shape of the delayed functions, and produced a worse WAIC value than the delayed-power model in all data sets, so we do discuss it further.

Results

Table 2 provides the WAIC values for each law for each data set, with the winning function's WAIC in bold for each data set. In terms of the preferred function, one of the two exponential laws was selected in 15 of the 17 data sets, with only

[1] For the exponential function the effect of pre-experimental learning can be accounted for by the value of β as the exponential has a translation-invariant shape.

two data sets showing a preference for a power law. In 13 of the 17 data sets WAIC selected functions with a delay parameter, with one each of the motor learning (K1), mental arithmetic (M3) and memory and visual search (S1 and V3) data sets constituting the four exceptions, which were all exponential. In 11 cases, coming from every task, including all three counting tasks, the delayed exponential was selected. The delayed power model was selected in two of the five mental arithmetic data sets. The standard power model was never selected.

In order to quantify the strength of these selections we divided the difference in WAIC between the best and second best model by the standard error of the difference. Standardized differences (which are analogous to a z-score) less than two provide only tentative evidence (Gelman, Carlin, et al., 2014). All of the four data sets that rejected a delay parameter in the exponential, the rejection was strong. In nine of the 11 selections favoring the inclusion of a delay in the exponential function were strong. The two cases only tentatively favoring a delay in the exponential were from mental arithmetic tasks. Of the two selections favoring the delayed power function, which were also both in mental arithmetic, one was strong relative to the standard exponential, and the other was very tentative relative to a delayed exponential.

Figure 3 illustrates fits to data from the motor learning (K1, upper) and alphabet arithmetic (A1, lower) experiments. In order to avoid the well known distortions associated with averaging, we only plot as examples data for a single subject and condition. The upper panel shows data from the motor learning experiment, where there was little evidence for an initial flat section, so that standard and delayed function fits are almost identical (only the standard versions are plotted in the figure).

The parameter for the exponential function (see figure caption) predict a median RT of 1.71s (i.e,. $t_0 + \alpha$) and standard deviation of 0.25s after extensive practice[2], with a lower bound for RT of 0.19s (i.e., t_0). In contrast, for the power function the corresponding estimates are implausibly small, 0.11s, 0.006s and 0.07s, respectively. At the other end of practice, the exponential function has a median RT on trial one of 6.3s, while for the power function it is 15.9s; the power estimate is clearly implausibly high.

The lower panel shows a subject from the alphabet arithmetic experiment, where there was clear evidence for a delay. Learning is of the type classified as a "smooth speed-up" by Rickard (2004; see his Figure 4). The parameters for the delayed exponential function predict a median RT of 0.84s and standard deviation of 0.23s after extensive practice, with a lower bound for RT of 0.43s. The corresponding estimates for the delayed power are 0.32s, 0.007s and 0.31s, with the standard deviation clearly being an underestimate, due to an implausibly small estimate the variable portion of asymptotic RT (0.01s). **Indeed, in both of the cases shown in Figure 3 the power functions displayed the tendency noted by Heathcote et al. (2000) to predict faster asymptotic performance than is plausible for empirical data.** The delayed exponential function has a median RT of 5.7s on the first trial, whereas the delayed power provides what appears to be an overestimate of 14.7s. Examination of Table 1 shows that the delayed exponential function provided plausible asymptotic estimates of the wide range of performance in the 17 data sets examined by Heathcote et al. (2000), ranging from the the fastest case in involving

[2]The asymptotic standard deviation is the product of the variable part of asymptotic median RT, $\alpha + \beta$, and the standard deviation of $LN(0, \sigma)$, which is $\sqrt{e^{\sigma^2} \times (e^{\sigma^2} - 1)}$.

memory search (S3: median = 0.35s, SD = 0.06s and lower bound = 0.13s) to the slowest involving the production of a complex sequence of key presses (K1: median = 2.32s, SD = 0.42s and lower bound = 0.93s). Similarly, estimates of median RT on the first trial ranged widely (e.g, 0.41s in S3 and 5.43s in K1).

Although much of the behavior in practice paradigms appears lawful it is inevitable that there will also be outlying data points. Such outliers are important to identify, both theoretically, because they may indicate the presence of a mixture of processes that underpin performance, and practically in cases where they distort parameter estimates and goodness-of-fit to the main body of the data. However, outliers can be hard to discern in the present context because of the non-linear effects of practice and the systematic decrease in variability with practice, so that deviations later in practice (when variability is lower) should be given more weight than variations earlier in practice.

Figure 4 show a method of summarizing goodness-of-fit and detecting outliers using a "QQ" (quantile-quantile) plot. This plot enables an easy determination of whether a model's characterization of RT data is accurate, not only in terms of its central tendency, but also the full distribution of RT. This compact graphical summary, which can be used with all of the practice functions considered here, is obtained by transforming the data for a participant and condition using the model's posterior median parameters. Other measures of the posterior central tendency, or even maximum likelihood estimates, could also be used. Firstly, the t_0 estimates are subtracted from the data, and the remainder divided by the values given by the term in brackets in Equation 6 (or Equation 5). Lastly, the natural logarithm

is taken, and the result divided by the lognormal standard deviation estimate. If the model is accurate, the results of these transformations should have a standard normal distribution.

To display the results graphically the sorted transformed data (i.e., its quantiles or "order statistics") are plotted as a function of the corresponding theoretical quantiles for a standard normal distribution[3]. This allows every data point to be shown on the same scale on a single QQ-plot (data from multiple conditions and/or participants can also be plotted together in this way as they are all also put on the same scale). If the fit is good the points on the QQ plot should fall on the main diagonal. Figure 4 shows QQ plots for the delayed exponential law for the sets in Figure 3 is indeed very good, lacking any systematic deviations from the main diagonal that would be indicative of a failure of its assumptions or distortion in the fit caused by influential outliers. For example, although the alphabet arithmetic QQ plot does bring attention to two potential positive outliers, corresponding to the two high points at trials 1749 and 1868 in the lower panel of Figure 3, they do not appear to have had much effect on the overall fit.

Lastly, Figure 5 provides an experiment-level assessment of the goodness-of-fit for the delayed exponential law. These plots were based on the transformed mean RT for nineteen evenly spaced trials for each condition for each subject. The transformed mean RTs for these semi-deciles were then averaged over subjects and conditions, forming a single set of

[3]The cumulative probability of the i^{th} of n sorted observations can be estimated as $p_i = i/(n+1)$, and corresponding standard normal quantile is $z(p_i)$, where z is the inverse standard cumulative normal transformation.

deciles for the entire experiment. We used trial semi-deciles in order to align practice series of different lengths before averaging. A good fit is indicated by the points in the plot clustering around the main diagonal, and, in general, the model predictions do appear to quite closely match the empirical data, suggesting that the delayed exponential provides a fairly accurate account of all datasets at the experiment-level. The fits to K1 and A1 datasets are better at the experimental-level than the individual conditions for the individual subjects displayed before, suggesting that some of the misfit in the latter was likely due to random variation rather than systematic misfit. As might be expected, fit is somewhat worse in datasets with very few trials (e.g., M3, M4, M5).

Fitting Sharp Transitions

None of the data examined by Heathcote et al. (2000) had sharp transitions of the type simulated in Figure 2. In contrast, they were evident in the majority (71%) of practice series in Rickard's (2004) data. For this data WAIC clearly favored the exponential functions, and selected the delayed exponential law (13359) over the standard exponential law (13965), and the delayed power law (14422) was selected over the standard power law (15818). In terms of standardized differences the evidence favoring the inclusion of a delay parameter in the exponential law (15.0) was the stronger than anything else previously reported, and was even stronger for the inclusion of a delay parameter in the power law (27.5). These results reinforce the inadequacy of the standard power and exponential functions, but a question remains as to whether the new delayed versions provide an adequate account. Answering this

question requires model checking rather than model selection.

Figure 6 displays the goodness of fit of the delayed exponential model to the data from a single subject in two cases displaying Rickard's (2004) Type 1 and Type 2 transitions. For the Type 2 transition on the right of Figure 6, analogous to the top right panel of Figure 2, most of the data lie along the main diagonal of the QQ plot, consistent with a good fit. This may be somewhat surprising given that there are some apparent slow outliers in the practice function plot, particularly trials 15 and 16 early in practice, trials 24 and 25 in the middle and trial 46 late in practice. The latter three values are indeed the largest positive values in the QQ plot, but the former two are only 4th and 6th largest, because they occur where RT variation would be expected to be much larger. Further, stronger positive deviations are to be expected given RT distribution is positively skewed, so few of these points lie far from the main diagonal. Although we do not claim that this fit is perfect, it does serve to illustrate the importance of judging outliers relative to a reasonable model RT variability and its modulation by practice.

In a Type 1 transition, shown on the left of Figure 6, there are strong slow outliers late in practice (a cluster of three on trials 29, 30 and 31 and one on trial 45) when the majority of performance is fast (analogous to the bottom right panel of Figure 2). Figure 6 also show a complimentary type of fast outlier that occurred early in practice (trial 11) when the majority of performance is slow. The QQ plot shows that the outliers (the four points on the upper right and one on the lower left) strongly distort the fits to the main part of the data, which is quite linear in the QQ plot but with a shallower slope, indicating that the level of variability (σ)

is over-estimated. In the practice function plot it is also evident that performance later in practice is over-estimated. This also results in the fast data point at trial 33 having the second largest negative deviation in the QQ plot. These shortcomings, particularly given the prevalence of Type 1 transitions (around half of all practice series), suggest the delayed random exponential practice law is not an adequate model for Rickard's (2004) data. As we now discuss, we believe that one of the most useful features of general laws (in this case not only the shape of the practice function but the nature of RT variability) is their ability to highlight exceptions requiring further investigation.

Discussion

In an influential review titled "Why the Laws of Memory Vanished" (Roediger, 2008b) argued that simple laws do not hold because of "the complex, interactive nature of memory phenomena" (p. 225). Although we agree about the complexity of learning and memory we believe it is this very complexity that makes valuable as a benchmark the widely applicable – if not perfect – regularities encapsulated by laws. This is particularly so when the laws are quantitative and comprehensive, in the sense of taking account of not only the central tendency but also the variability in behavior. For example, in the domain of response time measurement (RT) exceptions to unimodal positively skewed RT distribution call attention to the likely presence of a mixture of underlying cognitive processes. Similarly, in the domain of practice Heathcote et al. (2000) argued that exceptions to their finding for an exponential law likely signal the presence of a mixture of learning processes operating at different

time scales. Here we have brought together these pervasive regularities, along with a linear relationship between the mean and standard deviation of RT (Wagenmakers & Brown, 2007), and the presence of a slower period of learning early in practice, into a single law. We have presented evidence that this "Delayed Random Exponential Law of Practice" holds widely and provides a useful measurement model that distills complex patterns of behavior into few psychologically meaningful and practically useful parameters.

However, we acknowledge that outside of controlled experimental settings, and sometimes even within them, there may be exceptions: practice can slow rather than speed performance when fatigue sets in, performance can plateau when the potential for improvement in one method is exhausted, only to commence improving again when a new and better method is found, improvement can be sudden rather than gradual when insight is gained into a new much better method, but improvement can be inconsistent with occasional reversions back to the slower method. In the light of these facts, we have also sought to complement our refined law of practice with a sensitive method that detects when it does not hold. We contend that, against the background of a quantitative account of simple lawful behavior, exceptions – and the attendant need for deeper examination – become most evident. This can help motivate a need for the development of more complex theories.

One of our initial motivations of undertaking a re-analysis of the 17 data sets examined by of Heathcote et al. (2000), which are all well described by a smooth and monotonic decrease in RT with practice, was the possibility that two factors may have confounded their rejection of the Power Law of Practice. First, their analysis

did not take account for functional form complexity, so it is possible that the exponential function won only because it is more flexible than a power function having the same number of parameters. Second, neglecting an early period of slow learning would particularly disadvantage the power function, because it decreases more rapidly than the exponential function early in practice. Our re-analysis addressed both issues, using a Bayesian model-selection method that takes account of functional form complexity and by fitting practice functions that allowed for an early period of slow learning.

We found only one of the 17 data sets where the latter problem likely occurred, a mental arithmetic task (M1 in Table 2) where the standard exponential function beat the standard power function, but was clearly worse than a power function that allowed for a delay in the onset of more rapid learning. In all but one other case exponential functions were superior, and in most cases strongly so, with the only other exception, again in a mental arithmetic task (M4), being equivocal. Although rare, we do not wish to discount the exceptions; instead they highlight that our methods can find in favor of a power function and likely indicate the potential for learning in mental arithmetic to be multi-scale.

However, at a more general level, our refined analysis clearly indicates that Heathcote et al.'s (2000) repeal of the Power Law of Practice, and their proclamation of an Exponential Law of Practice, was justified. There would, then, seem to be no reason for the continued use of power functions to characterize practice effects. This is especially so as power functions demonstrably produce parameter estimates that provide an implausible characterization of both initial and asymptotic behavior as

well as misfit of the behavior in between. This makes them likely to be misleading in any practical application. Further, our results have the strong theoretical implication that practice effects usually have a single scale, suggesting a re-evaluation of theories that fundamentally assume a Power Law of Practice (e.g., Logan, 1988, 1992), and a re-formulation of those that assume it (e.g., J. R. Anderson & Lebiere, 1998)

Our results provided strong support for the addition of several extensions to the law of practice. First, there was a clear need to account for a delay in the onset of more rapid learning in 15 of the 18 data sets examined here. In the three exceptions there was strong evidence against the need to include a delay, so a delay is clearly not obligatory. However, the often strong evidence in favor of a delay in the remaining cases, as well as the single case where its omission may have caused a bias against detecting multi-scale learning, argues for the customary inclusion of a delay parameter when fitting practice functions.

A further addition, of a Lognormal RT distribution along with the assumption that it interacts multiplicatively with the practice function, was supported by the accurate characterization it provides of RT variability. These additions are not only necessary to provide a more complete description of behavior in practice paradigms; we would argue some assumptions about variability are unavoidable because they must be made, if only implicitly, when fitting practice data.[4] Instead, we think it is preferable to make explicit assumptions on this issue, and that to do so is especially

[4]**For example least squares fitting implicitly assumes a Gaussian error model. A reviewer suggested that in some cases this type of fitting might be preferable when interest focuses on the expected values. However, in a parameter recovery study we found that the τ and r parameters were less well estimated using least squares, and that this problem became more marked as RT variability increased.**

advantageous when, as was the case with our assumptions, they have attendant benefits: efficient estimation, rigorous model selection, and model checking via a graphical characterization of all aspects of fit in the from of easily constructed QQ plots.

This method of characterizing behavioral variability also underpins future potential developments in the quantitative characterization of practice effects, addressing limitations in the present approach. First, although we neglected the occurrence of errors in our analysis here (all RTs were from correctly answered trials), the Lognormal distribution is compatible with an evidence-accumulation approach to modeling not only the distribution of RT but also the probability of making different response choices (Heathcote & Love, 2012; Rouder et al., 2014; Terry et al., 2015). Although simultaneously modeling both types of data is challenging, as errors tend to be relatively rare especially later in practice, the progress that has already been made in very large data sets with relatively modest learning (Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Heathcote & Hayes, 2012) could be built upon by expanding the approach taken here.

Second, for practical applications more exploration is needed of how the delayed random exponential function can be used as a measurement model. Our analysis assumed one pair of population location and scale parameters for each model parameter in each condition in order to impose minimal assumptions. Better psychometric properties are likely to be obtained by a more parsimonious approach that shares scale parameters across conditions, and provides an account of within-subject correlations at the population level using an effect sizes parameterization. This could

also potentially support a more informed or perhaps default prior specifications, and the use of Bayes Factors for model selection as in (Rouder, Morey, Speckman, & Province, 2012).

Finally, we think it is important to acknowledge that practice data alone, and hence laws that are fit to those data, may not provide a full account of effects of practice. For example, recent evidence from brain imaging (Tenison & Anderson, 2016; Tenison, Fincham, & Anderson, 2016) suggests practice impacts progress through three stages in solving math problems, first characterized by a predominance of algorithmic processing, then retrieval and finally automatic processing. The first two stages might reasonably map to our delay and learning parameters respectively, and the third to our asymptote parameter. However, because performance is unchanging at asymptote, it is difficult investigate the nature of the automatic stage without alternative approaches, such as neuroimaging measures or behavioural manipulations such as imposing a dual-task load.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Earlbaum.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, *25*, 724–730.

Anglim, J., & Wynton, S. K. (2015). Hierarchical bayesian models of subtask learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 957–974.

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*, 25–35.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Brown, S., & Heathcote, A. (2003). Bias in exponential and power function fits due to noise: Comment on Myung, Kim, and Pitt. *Memory & Cognition*, *31*, 656–661.

Bryan, W. L., & Harter, N. (1897). Studies in the physiology and psychology of the telegraphic language. *Psychological Review*, *4*, 27–53.

Bryan, W. L., & Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review*, *6*, 345—375.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304.

Carrasco, M., Ponte, D., Rechea, C., & Sampedro, M. J. (1998). "Transient structures":

The effects of practice and distractor grouping on within-dimension conjunction searches. *Attention, Perception, & Psychophysics*, *60*, 1243–1258.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the stroop effect. *Psychological Review*, *97*, 332–361.

Crossman, E. R. F. W. (1959). A theory of the acquisition of speed–skill. *Ergonomics*, *2*, 153–166.

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, *16*, 1026–1036.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140.

Evans, N. J., & Brown, S. D. (2017a). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior Research Methods*, 1–15.

Evans, N. J., & Brown, S. D. (2017b). People adopt optimal policies in simple decision-making, after practice and guidance. *Psychonomic Bulletin & Review*, *24*, 597–606.

Evans, N. J., Howard, Z. L., Heathcote, A., & Brown, S. D. (2017). Model flexibility analysis does not measure the persuasiveness of a fit. *Psychological Review*, *124*, 339–345.

Evans, N. J., Rae, B., Bushmakin, M., Rubin, M., & Brown, S. D. (2017). Need for closure is associated with urgency in perceptual decision-making. *Memory & Cognition*, *45*, 1193–1205.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Aki, V., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed.). CRC Press.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information

criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207.

Heathcote, A., & Hayes, B. (2012). Diffusion versus linear ballistic accumulation: Different models for response time with different conclusions about psychological mechanisms? *Canadian Journal of Experimental Psychology*, *66*, 125–136.

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in psychology*, *3*.

Heathcote, A., & Mewhort, D. (1993). Representation and selection of relative position. *Journal of Experimental Psychology: Human Perception and Performance*, *19*, 488–516.

Josephs, R. A., Silvera, D. H., & Giesler, R. B. (1996). The learning curve as a metacognitive tool. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 510–524.

Keller, F. S. (1958). The phantom plateau. *Journal of the Experimental Analysis of Behavior*, *1*, 1–13.

Kramer, A. F., Strayer, D. L., & Buckely, J. (1990). Development and transfer of automatic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 505–522.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

Lee, F. J., & Anderson, J. R. (2001). Does learning a complex task have to be complex?: A study in learning decomposition. *Cognitive psychology*, *42*, 267–316.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

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, and Cognition*, *18*, 883–914.

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

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory and Cognition*, *28*, 832–840.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 324–354.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*, 559–586.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 435–451.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, *126*, 288–311.

Rickard, T. C. (2004). Strategy Execution in Cognitive Skill Learning: An Item-Level Test of Candidate Models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 65–82.

Rickard, T. C., & Bourne Jr, L. E. (1996). Some tests of an identical elements model of basic arithmetic skills. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1281-1295.

Roediger, H. L., III. (2008a). Relativity of remembering: Why the laws of memory vanished. *Annu. Rev. Psychol.*, *59*, 225–254.

Roediger, H. L., III. (2008b). Relativity of Remembering: Why the Laws of Memory Vanished. *Annual Review of Psychology*, *59*, 225–254.

Rosenbloom, P., & Newell, A. (1987). Learning by chunking: A production system model of practice. *Production system models of learning and development*, 221–286.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2014). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*, 1–23.

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 3–29.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Seibel, R. (1963). Discrimination reaction time for a 1,023-alternative task. *Journal of Experimental Psychology*, *66*, 215–226.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model

evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Snoddy, G. S. (1926). Learning and stability: A psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology*, *10*, 1–26.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583–639.

Strayer, D. L., & Kramer, A. F. (1994a). Aging and skill acquisition: Learning-performance distinctions. *Psychology and Aging*, *9*, 589–605.

Strayer, D. L., & Kramer, A. F. (1994b). Strategies and automaticity: I. Basic findings and conceptual framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 318–341.

Strayer, D. L., & Kramer, A. F. (1994c). Strategies and automaticity: II. Dynamic aspects of strategy adjustment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 342–365.

Suppes, P., Fletcher, J. D., & Zanotti, M. (1976). Models of individual trajectories in computer-assisted instruction for deaf students. *Journal of Educational Psychology*, *68*, 117–127.

Tenison, C., & Anderson, J. R. (2016). Modeling the distinct phases of skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 749–767.

Tenison, C., Fincham, J. M., & Anderson, J. R. (2016). Phases of learning: How skill acquisition impacts cognitive processing. *Cognitive psychology*, *87*, 1–28.

Terry, A., Marley, A., Barnwal, A., Wagenmakers, E.-J., Heathcote, A., & Brown, S. D. (2015). Generalising the drift rate distribution for linear ballistic accumulators.

Journal of Mathematical Psychology, *68*, 49–58.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*, 368–384.

Ulrich, R., & Miller, J. (1993). Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, *37*, 513–525.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*, 1413–1432.

Verwey, W. B. (1996). Buffer loading and chunking in sequential keypressing. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 544–562.

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*, 830–841.

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology*. New York: Holt.

Yarrow, K., Brown, P., & Krakauer, J. W. (2009). Inside the brain of an elite athlete: The neural processes that support high achievement in sports. *Nature Reviews Neuroscience*, *10*, 585–596.

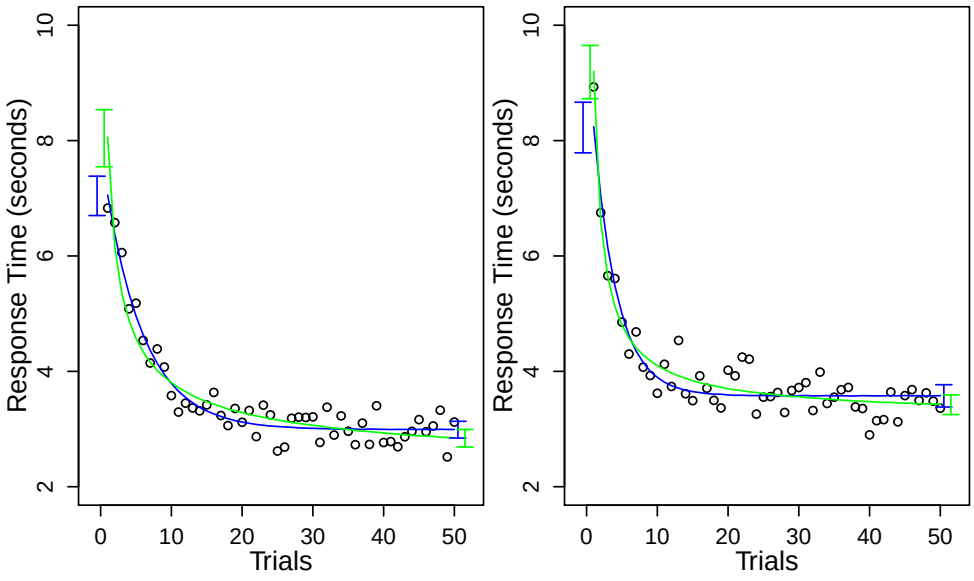

Figure 1. Fits to two simulated data sets (black circles), generated by the exponential function (left panel; parameters: $\alpha = 1$, $\beta = 2$, $r = 0.2$, $t_0 = 0.3$, and $\sigma = 0.1$) and the power function (right panel; parameters: $\alpha = 1$, $\beta = 2$, $r = 0.7$, $t_0 = 0.3$, and $\sigma = 0.1$). The exact exponential and power models used to simulate the data are the shifted multiplicative lognormal forms discussed later in the "Accounting for the RT Distributions" section. The lines are maximum likelihood estimated best fitting power and exponential functions, in green and blue respectively. In each graph, the x-axis represents the trial number, and the y-axis represents the response time on each trial. For the exponential data in the left panel the best-fitting power function overestimates response times in the middle, and underestimates them late in practice. For the power data in the right panel the best-fitting exponential function underestimates response times in the middle, and overestimates them late in practice. The error bars presented before the first trial provide the inter-quartile range of the data-generating model for the first trial, showing the predicted variability in the data at the start of practice. The error bars presented after the last trial provide the inter-quartile range for the last trial, showing both data-generating models decrease variability with practice.

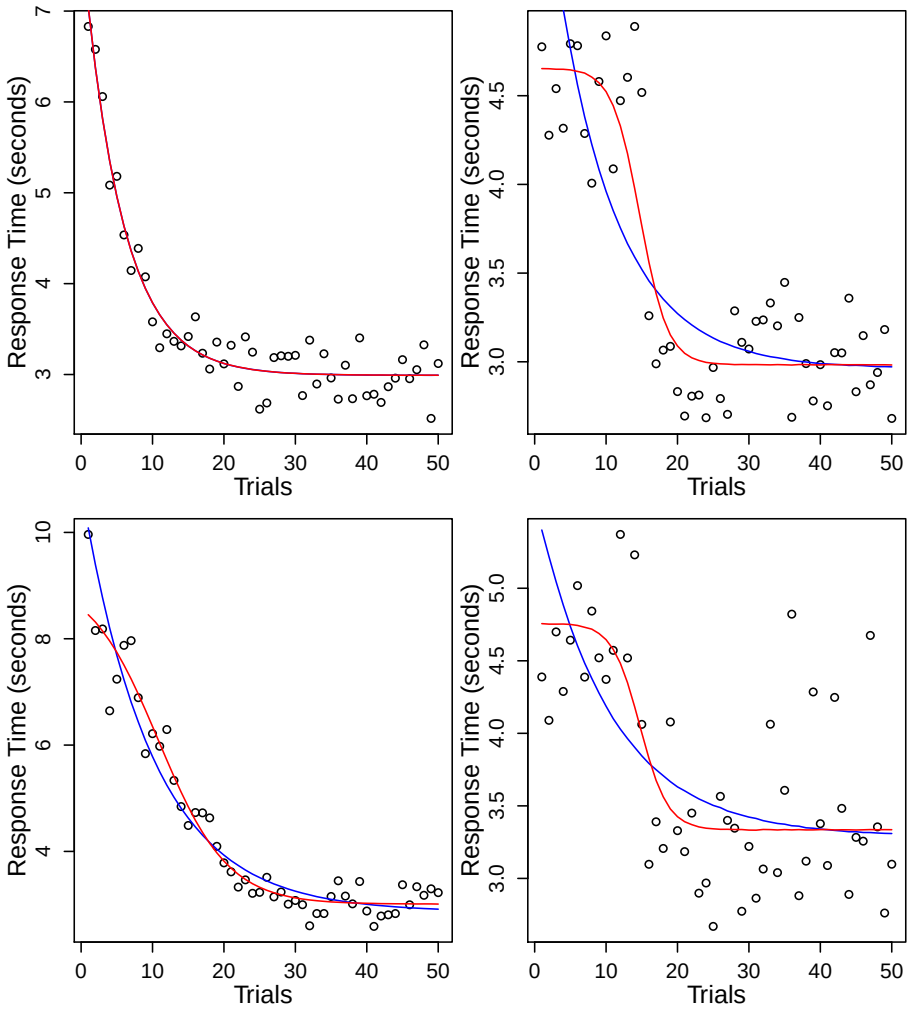

Figure 2. Maximum likelihood estimated best fits for the standard exponential (blue) and delayed exponential (red) models to four simulated data sets (black circles). Each data set was generated with the following common exponential parameters: $\alpha = 1$, $\beta = 2$, $r = 0.2$, $t_0 = 0.3$, and $\sigma = 0.1$ (see Equation 6). The data sets in the left column came from exponential (top left), and delayed exponential (bottom left, with $\tau = 10$) functions. The data sets on the top right came from a model with an initial period of no learning (the first 15 trials) generated from a lognormal distribution ($\mu = 1.5$, $\sigma = 0.1$), and the later trials generated by a standard exponential with multiplicative lognormal noise ($\sigma = 0.1$). The same model was used for the bottom right panel, except that trials after the 15th have a 0.2 probability of being samples from the same distribution as the initial 15 trials.

Table 1: Summary information about the data sets from Heathcote et al. (2000). The "Data set" column defines a name used to refer to the data set throughout this paper, and the "Source" column provides the corresponding key to the following citations: 1 = Strayer & Kramer (1994b), 2 = Strayer & Kramer (1994c), 3 = Strayer & Kramer (1994a), 4 = Palmeri (1997), 5 = Rickard & Bourne Jr (1996), 6 = Rickard (1997), 7 = Reder & Ritter (1992), 8 = Schunn et al. (1997), 9 = Heathcote & Mewhort (1993), 10 = Carrasco et al. (1998), 11 = Verwey (1996). The "Subs" column defines the number of subjects that are in the data set, and the "Conds" column the number of within-subjects conditions. All conditions were manipulated within subjects, except in M1 and V3 data sets, which also had three and two groups respectively (Note that M1 was manipulated in a non-factorial manner). The "Task" column refers to the type of task used in the data set, with the following key: 1 = Memory Search, 2 = Counting, 3 = Mental Arithmetic, 4 = Alphabetic Arithmetic, 5 = Visual Search, 6 = Motor Learning. The final 6 columns are the medians of the hyper-mean posteriors for the delayed exponential model, averaged over conditions.

Data set	Source	Task	Subs	Conds	τ	α	β	r	t_0	σ
S1	1,2	1	6	6	2.3	0.23	0.09	0.0012	0.25	0.41
S2	1,2	1	32	6	0.87	0.19	0.12	0.00086	0.24	0.35
S3	3	1	22	6	2.1	0.23	0.05	0.00065	0.13	0.26
C1	4	2	4	30	2.7	0.21	1.77	0.0019	0.4	0.41
C2	4	2	4	72	2	0.20	1.79	0.00084	0.43	0.44
C3	4	2	5	72	1.4	0.21	1.30	0.0007	0.44	0.41
M1	5	3	24	16	0.5	0.26	0.62	0.061	0.55	0.49
M2	6	3	19	12	0.9	0.28	7.44	0.0061	0.58	0.58
M3	7	3	20	4	2.8	0.47	4.53	0.16	0.14	0.85
M4	7	3	16	4	2.7	0.34	0.49	0.16	0.19	0.51
M5	8	3	22	3	3.7	0.28	0.63	0.11	0.20	0.65
A1	6	4	21	24	1.7	0.29	3.54	0.0035	0.46	0.52
V1	9	5	24	8	2.3	0.21	0.44	0.0035	0.31	0.49
V2	10	5	10	12	3.5	0.33	0.25	0.011	0.22	0.36
V3	9	5	8	16	1.6	0.25	0.23	0.0084	0.35	0.50
K1	11	6	36	2	0.5	1.80	3.14	0.0097	0.52	0.23
K2	11	6	36	2	4.2	0.40	0.83	0.0014	0.93	0.37

Table 2: WAIC values for each data set, where smaller values indicate a superior model, and the number in **bold** indicates the winning model for each data set. The final column is the standardized difference between the best and second best models.

Dataset	SE: Standard Exponential	DE: Delayed Exponential	SP: Standard Power	DP: Delayed Power	z(Best-Next)
k1	**20742**	20759	22642	20854	(SE > DE) 3.0
k2	-667	**-968**	31	-23	(DE > SE) 14.0
m1	-19779	-19585	-19679	**-19974**	(DP > SE) 4.3
m2	44051	**44021**	49205	46419	(DE > SE) 1.1
m3	**3164**	3176	3209	3195	(SE > DE) 2.2
m4	-411	-414	-403	**-424**	(DP > DE) 0.8
m5	4431	**4414**	4583	4499	(DE > SE) 1.5
s1	**-36868**	-36836	-36721	-36760	(SE > DE) 3.9
s2	-310045	**-310132**	-308662	-309522	(DE > SE) 3.7
s3	-287100	**-287174**	-286537	-286798	(DE > SE) 2.7
v1	-28228	**-28380**	-24868	-27019	(DE > SE) 5.7
v2	-11177	**-11232**	-10660	-10880	(DE > SE) 3.3
v3	**-29065**	-28947	-27702	-28381	(SE > DE) 7.0
a1	45602	**45186**	55228	50665	(DE > SE) 8.4
c1	-14358	**-15053**	-7712	-9886	(DE > SE) 13.1
c2	-24426	**-24937**	-13764	-16574	(DE > SE) 8.4
c3	-41478	**-41705**	-29766	-34099	(DE > SE) 3.9

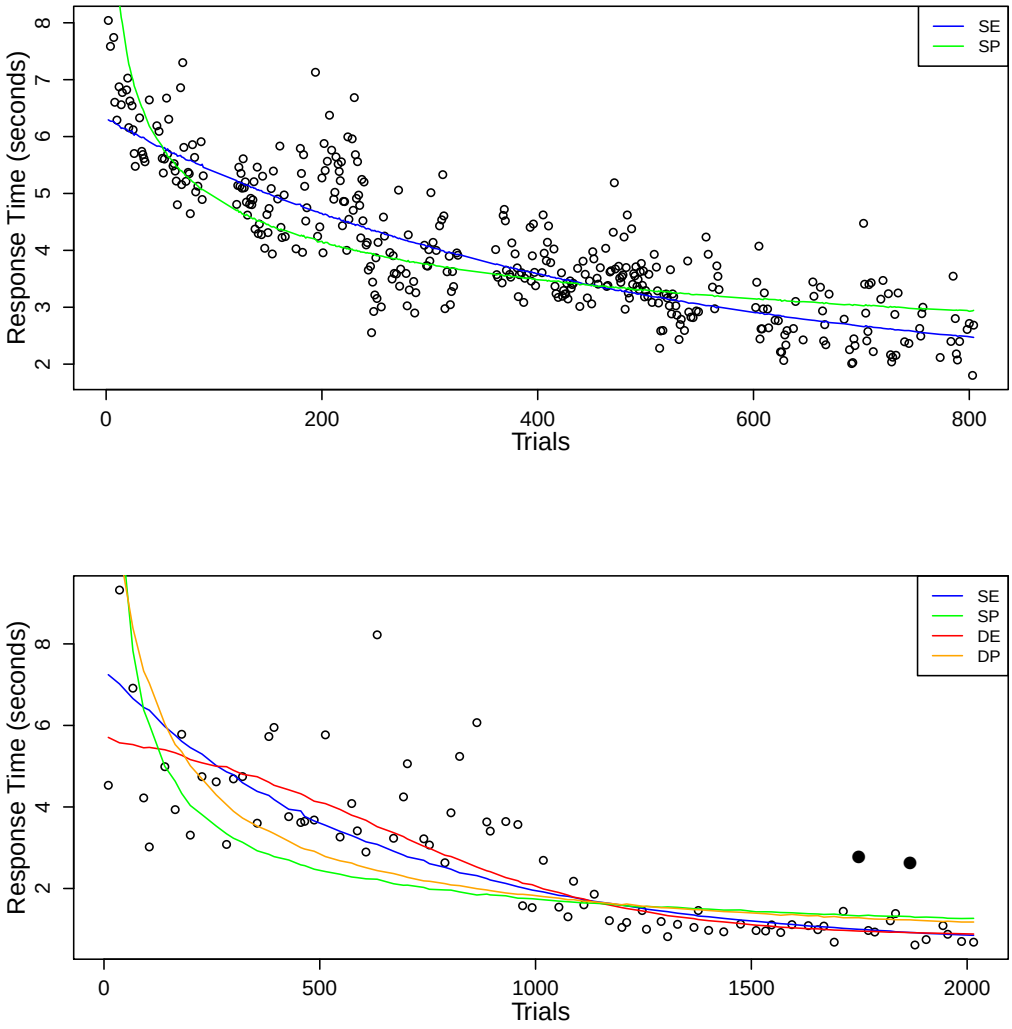

Figure 3. The upper panel displays data (black circles) and posterior median parameter fits of the median standard random exponential (SE; blue line: $t_0 = 0.19$, $\alpha = 1.52$, $\beta = 4.6$, $r = 0.0022$, and $\sigma = 0.16$) and power (SP; green line: $t_0 = 0.066$, $\alpha = 0.035$, $\beta = 15.8$, $r = 0.26$, and $\sigma = 0.18$) functions to condition 1 for subject 9 in the K1 data set. The lower panel displays same type of fits of the standard exponential (SE; blue line: $t_0 = 0.42$, $\alpha = 0.155$, $\beta = 6.8$, $r = 0.0016$, and $\sigma = 0.5$) and power (SP; green line: $t_0 = 0.4$, $\alpha = 0.014$, $\beta = 106$, $r = 0.63$, and $\sigma = 0.67$) and delayed exponential (DE; red line: $t_0 = 0.43$, $\alpha = 0.41$, $\beta = 4.9$, $r = 0.0033$, $\sigma = 0.48$, and $\tau = 8.2$) and delayed power (DP; orange line: $t_0 = 0.31$, $\alpha = 0.01$, $\beta = 14.4$, $r = 0.88$, $\sigma = 0.56$, and $\tau = 50$), to data from condition 24 of subject 21 in Experiment A1. The two filled black dots are the outlier data points discussed in the main text, in reference to Figure 4: trials 1749 and 1868

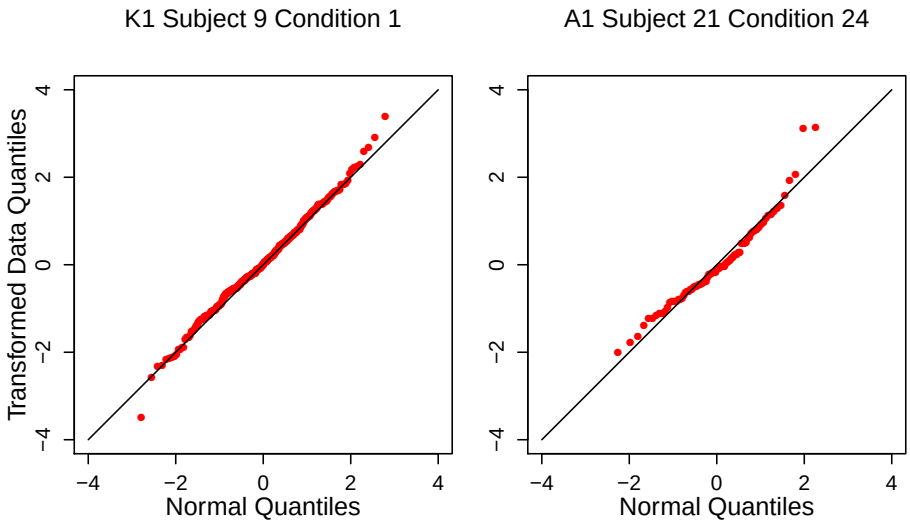

Figure 4. Quantile-quantile plots of residuals from the delayed exponential model for the data displayed in Figure 3. The X axis displays the quantiles under the standard normal, and the Y axis displays the sorted data transformed as described in the main text.

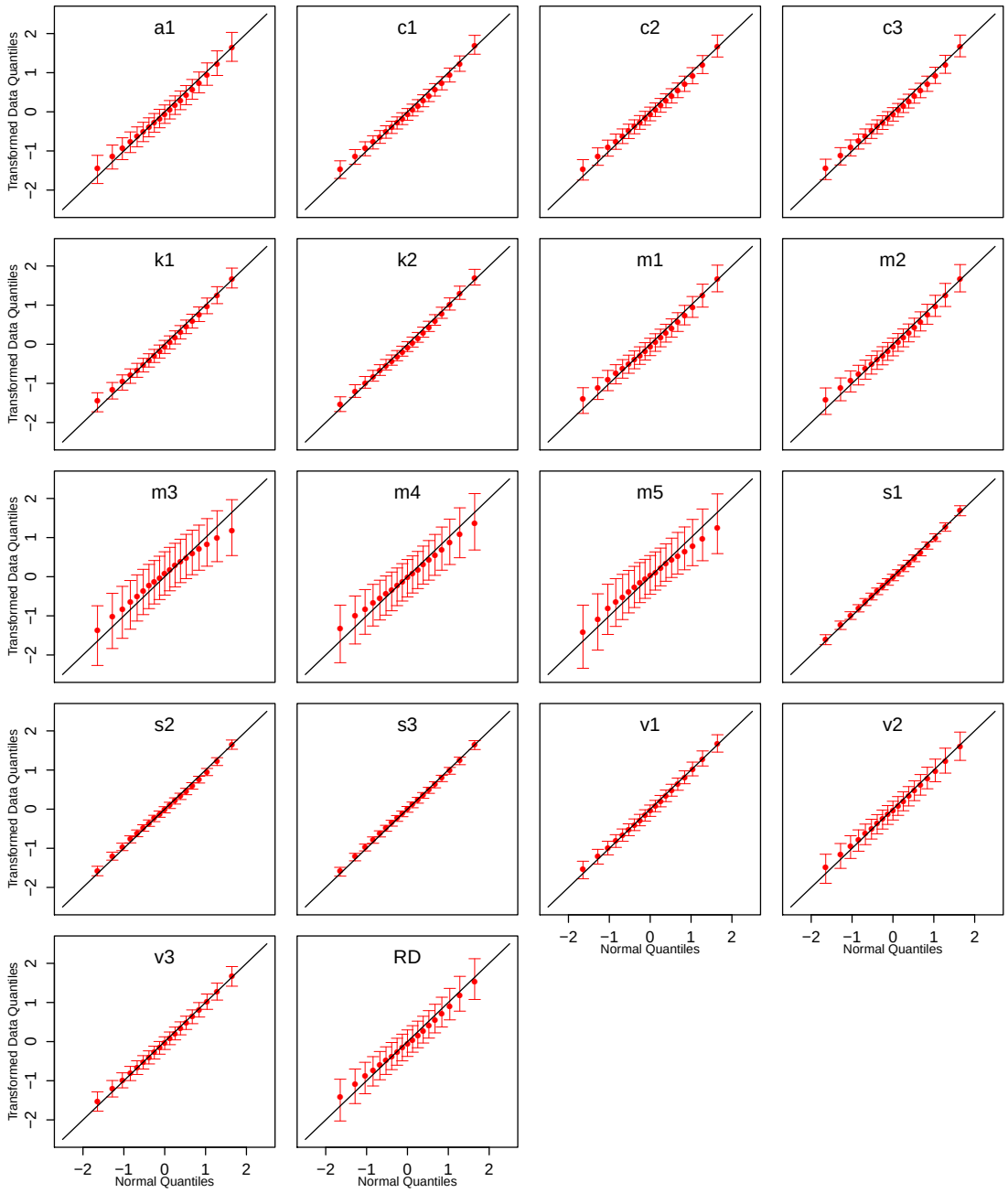

Figure 5. **Quantile-quantile plots of residuals from the delayed exponential model for the group-averaged semi-deciles trials for each experiment, with different points for each experimental condition. Error bars provide the 95% posterior credible intervals, which were generated by transforming the data generated by 100 different posterior samples, and then for each semi-decile, taking the 0.025 and 0.975 quantiles over the posterior samples. The X axis displays the quantiles under the standard normal, and the Y axis displays the sorted data transformed as described in the main text.**

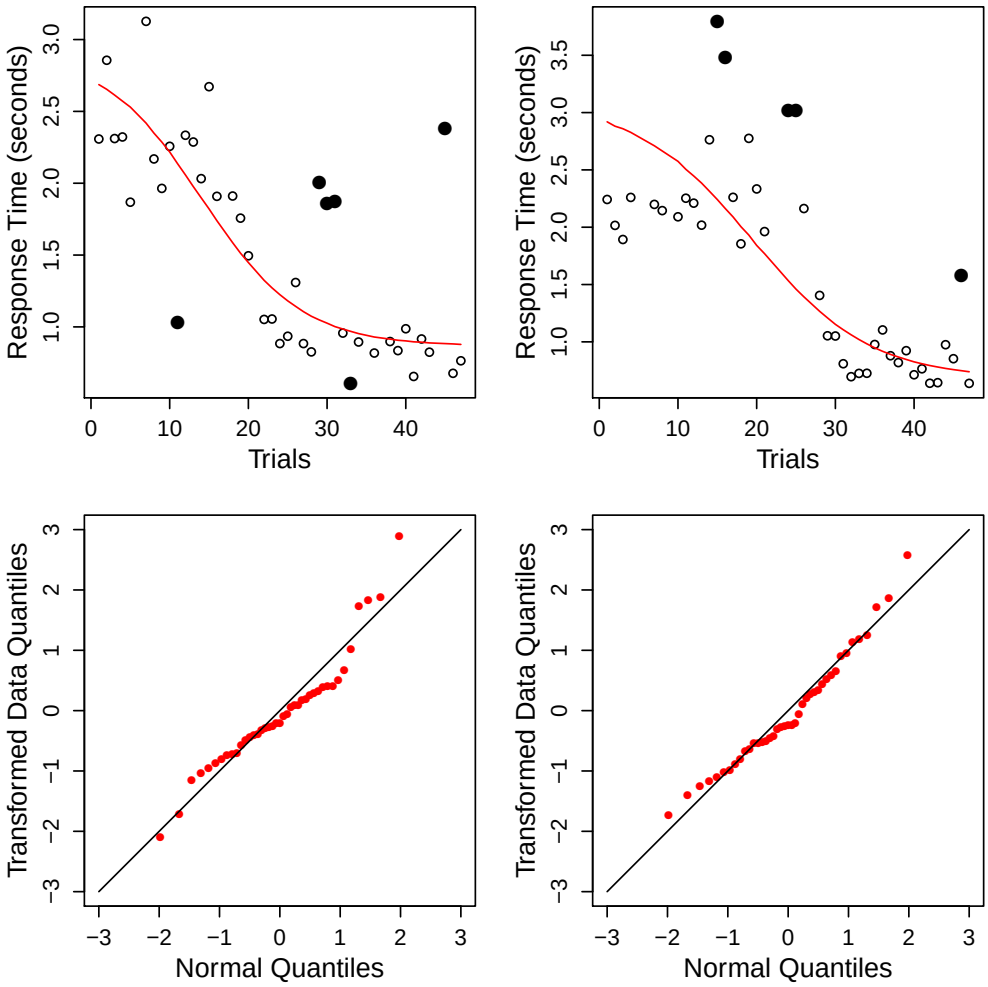

Figure 6. Practice series (upper panels) and corresponding QQ plots (lower panels) for subject 13 of Rickard (2004). The data in the left column is for an item displaying at Type 1 transition (slow late outliers and a fast earlier outlier) and a Type 2 transition, where such outliers are less prevalent. The red lines in the top panels show the fit of the posterior median of the median of the delayed random exponential function. In the top panels, the filled black dots are the outlier data points discussed in the main text, in reference to the bottom panels.