LEARNING CURVES AS STRONG EVIDENCE FOR TESTING MODELS:
THE CASE OF EBRW

Denis Cousineau         Guy L. Lacroix
Université d'Ottawa    Carleton University

Gyslain Giguère         Sébastien Hédie
Université de Montréal  Purdue University

Running head: EBRW and Learning curves

For correspondence:

Denis Cousineau
École de psychologie
Université d'Ottawa
136, Jean-Jacques Lussier
Ottawa (Ontario) K1N 6N5 CANADA

Office: (613) 562-5800 #7910
E-mail: denis.cousineau@uottawa.ca
Authors’ Note

Denis Cousineau, École de psychologie, Université d’Ottawa; Guy L. Lacroix, Department of Psychology, Carleton University; Gyslain Giguère, Département de psychologie, Université de Montréal; Sébastien Hélie, Department of Psychological Sciences, Purdue University.

This research was supported in part by grants from the Conseil Canadien pour la Recherche en Sciences Naturelles et en Génie awarded to the first and second author, and by fellowships from the Fonds Québécois de Recherche sur la Nature et les Technologies awarded to the third and fourth author. We thank Nick Chater, Andrew Heathcote, Gordon Logan, Robert Nosofsky and Emmanuel Pothos for helpful comments on a previous version of this manuscript. We also thank Glen Howell for running the replication of the original Palmeri and Nosofsky (1997a) study 1.
Abstract

In this manuscript, we show how learning curves can be used to provide a strong test for computational models of cognitive processes. As an example, we show how this method can be used to evaluate the Exemplar-Based Random Walk model of categorization (EBRW; Nosofsky & Palmeri, 1997a). EBRW is an extension of the Generalized Context Model (GCM; Nosofsky, 1984, 1986). It predicts that mean RTs follow a power function. It can be shown analytically, however, that the learning rate (i.e., the curvature) predicted by the model can only be equal to 1, a value rarely observed in empirical data analyses. We also explored an extended version of EBRW including background noise elements (Nosofsky and Alfonso-Reese, 1999) and identified conditions under which this model can predict curvatures different from 1. This shortcoming can be resolved by a simple extension to EBRW in which the original exponential distribution of retrieval times is replaced by a Weibull distribution. Additional predictions regarding learning curves are discussed.

Keywords

Learning curves; power curve; categorization models; Exemplar-based random walk model.
Learning curves as strong evidence for testing models:

The case of EBRW

Learning curves have been used to evaluate knowledge and skill acquisitions across time since the beginnings of experimental psychology (Ebbinghaus, 1885/1962; Thorndike, 1911). Much of the theoretical work on learning curves has focused on describing them mathematically as a power function (also called the power curve; Newell & Rosenbloom, 1981), an exponential function (Heathcote, Brown & Mewhort, 2000), a mixture of both (Anderson & Tweney, 1997), or other functions (e.g. Rickard, 1997; Wixted & Ebbesen, 1997; Newell, Mayer-Kress, & Liu, 2006). One aspect of learning curves which in our opinion has been relatively neglected is their value as a diagnostic tool for supporting or rejecting models of learning. While computational models are usually fit to some measure of central tendency for the dependent variable(s) (e.g., the mean), fitting the learning curve trajectory is a much more stringent test. To make this case, we propose to examine the Exemplar-Based Random-Walk model’s (EBRW: Cohen & Nosofsky, 2000, 2003; Nosofsky, 1997; Nosofsky & Alfonso-Reese, 1999; Nosofsky & Palmeri, 1997a, 1997b, 2008; Nosofsky & Stanton, 2005, 2006; Palmeri, 1997a, 1997b, 1999) predicted learning curve and establish if it can be used to test the model’s validity in light of the available empirical data. The EBRW model was selected because it is one of the few models that is well-specified enough to analytically generate a predicted learning curve.

EBRW is a model of speeded classification behavior. Its kernel is the Generalized Context Model (GCM: Nosofsky, 1984, 1986; Shin & Nosofsky, 1992), whose outputs are used to drive a random walk inspired by Logan’s (1988; 1990) Instance-based Theory of
EBRW and Learning curves

Automaticity. EBRW is one of the few categorization models able to predict both response selection probabilities and latencies separately or simultaneously (see also Ashby, Boynton, & Lee, 1994; Cousineau, 2004; Kent & Lamberts, 2005; Lamberts, 1998, 2000). More importantly, it is one of the few models to provide predictions about the change in response times (RTs) with practice.

The goal of the present paper is to examine EBRW’s validity as a model of speeded classification, by looking at its predicted learning curves. First, we show that EBRW predicts learning curves for response times described by a power function (Newell & Rosenbloom, 1981; Rosenbloom, 2006), but with a specific curvature parameter \( c = 1 \). However, at least some empirical data can be shown to be best accounted for by a range of different learning rates. To show that EBRW’s inability to produce the range of curvatures observed in data is not due to artifacts produced by the fitting procedure, we collected new data and used Monte Carlo simulations. Thirdly, we examine a version of EBRW, called the Background-Noise Element EBRW and show that this version still cannot account for the whole range of observed learning curvatures. Fourthly, we consider another version of EBRW, the Weibull-distribution EBRW, which possesses the latitude to predict any curvature. In the conclusion, we highlight some of Nosofsky and Palmeri’s (1997a) predictions regarding learning curves that, if tested, could lead to the development of future models or the refinement of existing models of speeded classification.

1. A formal description of EBRW and an analysis of the curvature parameter

Nosofsky and Palmeri (1997a) suggest that the evolution of RTs in EBRW may follow a power function (Cousineau, Hélie, & Lefebvre, 2003; Lacroix & Cousineau, 2006; Newell &
EBRW and Learning curves

Rosenbloom, 1981). This curve expresses response times RT as a function of learning block N, such that

\[ RT(N) = bN^{-c} + a \]  

(1)

in which a represents asymptotic performance, b is the difference between the initial performance and the asymptote (i.e., improvement), and c is the curvature parameter which indicates the rate at which participants acquire new knowledge or assimilate a new proficiency (Cousineau & Lacroix, 2006; Heathcote, Brown & Mewhort, 2000). A value close to zero for parameter c indicates a very slow acquisition rate, whereas a large value shows a fast one.

In order to predict the RTs for a particular stimulus, one must derive the expected decision time for the model following the presentation of item i, \( E(T \mid i) \). In a random walk model such as EBRW, the expected time to categorize an item is found by multiplying the expected time to achieve a step towards one of the boundaries, \( E(T_s \mid i) \) by the expected number of steps before one of the response decision boundaries is reached, \( E(N \mid i) \): \(^1\)

\[ E(T \mid i) = E(T_s \mid i) \times E(N \mid i) \]  

(2)

Regarding the second factor, the expected number of steps is a classical ruin problem (Busemeyer, 1982; Feller, 1957). Assuming that all steps are independent, this quantity is equal to:

\[ E(N \mid i) = \frac{1}{p_i - q_i} \left( \frac{\theta_i^{A+B} + 1}{\theta_i^{A+B} - 1} (A + B) - \frac{\theta_i^A + 1}{\theta_i^A - 1} A \right) \]  

(3)

where \( A \) and \( -B \) represent decision boundary positions (in Nosofsky & Palmeri, 1997a, \( A = B \)), \( p_i \) and \( q_i \) are the respective probabilities given stimulus \( i \) of approaching boundaries \( A \) and \( -B \)
respectively \( (p_i + q_i = 1) \), and \( \theta_i = p_i / q_i \) represents the odds of approaching boundary \( A \) relative to approaching boundary \( -B \).

EBRW’s kernel, the GCM component (Nosofsky, 1984, 1986), computes \( p_i, q_i \) respectively, by summing the similarities between test item \( i \) and all members of one category in memory at block \( N \), relative to the summed similarities between test item \( i \) and all the members of category \( A \) and \( B \) at block \( N \). Hence, if there are two categories:

\[
p_i = \frac{S_{iA_N}}{S_{iA_N} + S_{iB_N}} \quad \text{and} \quad q_i = \frac{S_{iB_N}}{S_{iA_N} + S_{iB_N}}
\]

where \( S_{iA_N} \) and \( S_{iB_N} \) respectively represent the summed similarity between item \( i \) and all the memorized exemplars of categories \( A \) and \( B \) at block \( N \), so that the odds ratio becomes

\[
\theta_i = \frac{S_{iA_N}}{S_{iB_N}}
\]

It follows from equations 4 and 5 that \( p_i, q_i, \) and \( \theta_i \) depend on the content of the exemplar memory. Under controlled experimental manipulation, the content of the exemplar memory directly depends on the training block \( N \).

Regarding the first term of Eq. 2, an assumption concerning this variable’s distribution is necessary in order to specify the mean time to achieve a step (Cousineau, Brown, & Heathcote, 2004). Nosofsky and Palmeri (1997a) chose a shifted exponential distribution, which is mathematically convenient and a reasonable choice because it may constitute a good approximation to central processing times (Hockley, 1984). In EBRW, it is assumed that the retrieval rate for exemplar \( j \) from memory is given by the similarity of \( j \) to the test item \( i \), noted \( s_{ij} \). Under this condition, the time required to make a step is equal to that of retrieving the first exemplar from a pool of exemplars (i.e., the minimum time: Cousineau, Goodman, & Shiffrin,
2002). Under the exponential distribution, that time is given by the reciprocal of the summed similarities to all the memorized exemplars. Therefore, an exemplar is retrieved with expected time:

$$E(T \mid i) = \frac{1}{\sum_{j \in J_N} s_{ij}} + \alpha$$  \hspace{1cm} (6)$$

where $\alpha$ is an arbitrary shift in mean time and $J_N$ is the content of the exemplar memory at block $N$. If all the exemplars belong to mutually exclusive categories (A and B), the summed similarity can be subdivided into two sums, $\sum_{j \in J_N} s_{ij} = \sum_{j \in A_N} s_{ij} + \sum_{j \in B_N} s_{ij}$, so that Eq. 6 can be rewritten as:

$$E(T \mid i) = \frac{1}{S_{iA_N} + S_{iB_N}} + \alpha$$  \hspace{1cm} (7)$$

in which $S_{iA_N} = \sum_{j \in A_N} s_{ij}$ is the summed similarity of test item $i$ to all the memorized exemplars belonging to category A at block $N$, and $S_{iB_N} = \sum_{j \in B_N} s_{ij}$ is the equivalent for category B.

Nosofsky and Palmeri (1997a, 1997b) originally made the simplifying assumption that EBRW makes no encoding or retrieval errors. Therefore, EBRW’s memory content following a first block of training (hereby noted $J_1$) contains one copy of each of the $j$ exemplars, $J_1 = \{E_1, \ldots E_j\}$ in which $E_j$ is the $j^{th}$ exemplar. During the second block, new copies of all $j$ stimuli are stored once again, so that $J_2 = \{E_1, \ldots E_j, E_1, \ldots E_j\}$ is the concatenation of a set that contains two instances of exemplars $E_1$ to $E_j$. This process is repeated up to $J_N$. Moreover, the absence of noise during the encoding and retrieval processes makes any test item $i$ as similar to the first training exemplar encoded as to any subsequent “copy” that is later encoded. Consequently,
EBRW and Learning curves

\[ S_{iA_N} = \sum_{j \in A_N} s_{ij} \]

that is, the summed similarity to a category increases linearly with the number of repeated

exemplars in the memory content.

Going back to Eq. 4, we therefore have that

\[ p_i = \frac{S_{iA_N}}{S_{iA_N} + S_{iB_N}} = \frac{N \times S_{iA_1}}{N (S_{iA_1} + S_{iB_1})} = \frac{S_{iA_1}}{S_{iA_1} + S_{iB_1}} \] (9)

is independent of the block number. The same occurs for \( q_i \). Likewise,

\[ \theta_i = \frac{S_{iA_N}}{S_{iB_N}} = \frac{N \times S_{iA_1}}{N \times S_{iB_1}} = \frac{S_{iA_1}}{S_{iB_1}} \] (10)

is a constant independent of repeated exposure to the same stimuli. Hence, training does not

affect \( p_i, q_i \) and, consequently, \( \theta_i \). If the boundaries are held constant (as is assumed), then the

expected number of steps to reach a boundary does not change with the addition of copies of the

training exemplars acquired during training. The expected number of steps is therefore a

constant:

\[ E(N \mid i) = k_i \] (11)

At first glance, this result may seem counter-intuitive. Nevertheless, because \( p_i \) is relative
to the total content of the exemplar memory, the duplication of that memory’s content \( N \) times
does not change the relative importance of one category relative to the other. Hence, the

probability of taking a step toward one of the two categorical boundaries is always the same for a
given test stimulus, irrespective of the amount of training. Consequently, the second term of Eq. 2, \( E(N \mid i) \), is not responsible for the decrease in response times with training.

For the first term, \( E(T_s \mid i) \), by (6) we have

\[
E(T_s \mid i) = \frac{1}{S_{iA_n} + S_{iB_n}} + \alpha
\]

\[
= \frac{1}{N (S_{iA_1} + S_{iB_1})} + \alpha
\]

\[
= k_2 N^{-1} + \alpha
\]

(12)

for which \( k_2 = \frac{1}{S_{iA_1} + S_{iB_1}} \) is once again a constant independent of training.

Finally, if we combine Eq. 11 and Eq. 12, we get a solution to Eq. 2:

\[
E(T \mid i) = E(T_s \mid i) \times E(N \mid i)
\]

\[
= (k_2 N^{-1} + \alpha) \times k_1
\]

(13)

Following Nosofsky and Palmeri (1997a), we insert two scaling parameters in Eq. 13, \( k_3 \) and \( \mu_R \), so that the decision time at block \( N \) is transformed into a response time:

\[
RT(N \mid i) = k_3 E(T \mid i) + \mu_R
\]

\[
= k_3 \times k_2 \times k_1 N^{-1} + k_3 \times k_1 \times \alpha + \mu_R
\]

\[
= b N^{-1} + \alpha
\]

(14)

in which \( b = k_3 \times k_2 \times k_1 \) and \( a = k_3 \times k_1 \times \alpha + \mu_R \) where \( \mu_R \) is a residual processing time and \( k_3 \) scales arbitrary processing times into milliseconds. Eq. 14 is a power function of the form given in Eq. 1 in which the exponent \( c \) is restricted to have the value 1. It follows that EBRW can only predict a learning power function with one curvature parameter, that is \( c = 1 \).

2- Curvature parameters of empirical and simulated data

A foreseeable consequence of EBRW’s inability to produce a curvature parameter \( c \) different from 1 is that its fit to empirical data (when analyzed in terms of blocked response
times from separate individuals) will deteriorate as empirically observed $c$ values progressively become different from 1. To explore this prediction, the results of Nosofsky and Palmeri’s (1997a) Experiment 1 were examined. In this experiment, three participants were asked to classify 12 color stimuli built from two dimensions (brightness and saturation) into two categories. Participants received typical feedback learning instructions and classified each stimulus 150 times (once per block, 30 blocks per day for 5 days). The authors also obtained similarity judgments from the same participants, in order to derive a multidimensional space for the stimuli. This is standard procedure for the GCM component of the model (see Shin & Nosofsky, 1992). The collected data were subsequently used to test EBRW.

A visual inspection of Nosofsky and Palmeri’s (1997a) fits to empirical categorization response times averaged over blocks (Figure 5 and Figure 7, reproduced in the present Figure 1, first column) provides preliminary evidence for the prediction that EBRW will not fit data adequately when $c \neq 1$. Indeed, participant 2, whose curvature parameter differed the most from 1 ($c = 0.214$), led to the poorest EBRW fit ($r = .78$). Moreover, this fit was worse than that obtained with the power function model ($r = .86$). In comparison, participants 1 and 3’s curvature parameters were much closer to 1 ($c = 1.02$ for participant 1 and $c = 0.790$ for participant 3).

Consequently, EBRW’s fits to those data were much better ($r = .94$ and $r = .96$) and much closer to those predicted by the power function model ($r = .94$ and $r = .97$ respectively).

Because the sample of curvature estimates in Nosofsky and Palmeri (1997a) was very small (3 individuals only), we conducted a replication of Palmeri and Nosofsky’s (1997a) Experiment 1 which included 12 participants. The top part of Figure 2 shows the evolution of accuracy over the 150 blocks obtained during the five days of training averaged across...
participants. The mean response times per block were computed and the curvatures were estimated. The bottom part of Figure 2 shows the distribution of the estimated $c$ exponents for the sample whereas Figure 3 shows individual learning curves for all participants (ordered by their estimated exponents and sorted into slow, moderately fast, and fast learners). The average $c$ was below 0.3 with a small standard error ($0.27 \pm 0.06$ standard error of the mean). This result indicates that participants such as Nosofsky and Palmeri’s participant 2 are not atypical.

So far, the individual curvatures obtained from our categorization data seem to differ from EBRW’s prediction that $c$ will always be equal to 1. However, it is possible that the curvature estimation process is biased downward so that the estimate is smaller than 1. It is also possible that internal fluctuations (e.g., the order of presentation of the stimuli) in EBRW produce curvatures that look smaller than 1, despite the formal analysis provided in the first section. Either issue may be generating variability around this ideal curvature of 1.

To have a sense of this potential variability, we simulated EBRW. The simulations included each step of the random walk until one of the boundaries was reached. For each simulation, a random two-dimensional psychological space was constructed with 12 exemplars. EBRW’s parameters were also randomly generated, so that a large, and hopefully representative, set of variations in the learning curve could be explored ($c$, the sensitivity parameter of the GCM in EBRW’s kernel, was random uniform in the range $[0.5, 3.0]$, $\alpha$ was random uniform in the range $[0.1, 0.5]$, $w_1$ was random uniform in the range $[0.1, 0.9]$, and $A$ and $B$ were random integers in the range $[3, 8]$, not necessarily equal). Finally, because scaling parameters $\mu_R$ and $k_3$ have no influence on the curvature, they were set to 100 ms and 400 ms respectively. The model was trained for 150 blocks and the mean response time per block computed. Those mean RTs
were analyzed to get an estimate of the best-fitting power function’s exponent (Cousineau & Larochelle, 1997). The simulation was repeated 10,000 times.

The mean curvature parameter across the simulations was 0.998 with a standard error of 0.003 (standard deviation of 0.269).³ The smallest exponent observed was 0.195 whereas the largest was 2.277. Figure 4 shows the distribution of the curvature parameter across simulations. The vast majority of the simulations produced a learning curve with a curvature close to 1. This indicates that the fitting process did not introduce a systematic bias. Moreover, it provides additional validity to the demonstration given in the first section. Overall, 99.99% of the simulations produced a curvature parameter that was larger than .213, the parameter of participant 2 in Nosofsky and Palmeri (1997a). Clearly, that participant would have to be declared “unfit” by EBRW (Roberts & Pashler, 2000).

In sum, the results of the previous and the present section indicate that curvatures are often much smaller than 1 in human data. This observation cannot be explained by sampling error in the model or biases in the estimation process, as the simulations showed no such difficulties.

Determining the appropriate time scale in EBRW

One difficulty when fitting learning curves concerns the choice of the time scale (also noted by Verguts & Storms, 2004): should we use training session as a variable, training block or trial number? In Nosofsky and Palmeri’s (1997a) Experiment 1, each stimulus was presented once per block. Hence, EBRW’s exemplar memory encodes each exemplar once per block (this constitutes EBRW’s obligatory encoding assumption; Logan, 1988, proposed a similar assumption). As a consequence, the speedup in response times should occur as a function of the
block number $N$. Therefore, $N$ is the correct independent variable and should be used for fitting response times. In Nosofsky and Palmeri (1997a), there were 150 blocks: the learning curve should be measured with $N$ varying between 1 and 150.

It is crucial to correctly identify the independent variable when drawing a learning curve because the power function (Eq. 1), unlike the exponential, is a scale-dependent function (Cousineau et al., 2003). Hence, using an alternate time scale which averages blocks of trials stretches the curve and artificially inflates the learning parameter value ($c$). Brown and Heathcote (2003) raised similar concerns regarding between-subject averaging.

Nosofsky and Palmeri (1997a) showed in their Figure 5 (reproduced here in Figure 1, first column) the best-fitting power curves based on the data from the figure.

Because the best-fitting curve reported in Nosofsky and Palmeri (1997a) used an inadequate time scale assuming the model, we refitted their data using an $N$ going up to 150 (basically, the first point became point 2.5, the second became point 7.5, the last one became 147.5, etc. using the middle block of each chunk of 5 blocks). The best-fitting curves are shown in the right column of Figure 1. As can be seen, the quality of the fit to the 150-block data is as good as that of Nosofsky and Palmeri’s fits to the 5-block time scale. However, once the model-relevant scale is used, the curvature parameters become smaller (going on average from 0.67 to 0.45, a reduction of about one third), and thus further away from 1. With the model-relevant time scale, we see that all three participants (not only the second one) now have learning curves that are problematic for EBRW.
3- The Background Noise Element version of EBRW

The version of EBRW that we have discussed so far is the core version of the model. It was selected for a formal analysis because it is the most straightforward implementation of Nosofsky and Palmeri’s (1997a) core idea that categorization RTs may be explained by a model that combines the GCM and the Instance-based Theory of Automaticity. This version has the advantage of being mathematically tractable. However, like any model, it makes simplifying assumptions. The authors noted that the model starts with an empty memory system. This assumption, the authors noted, explains that accuracies are at ceiling right from the start (and therefore constant, as shown previously). To add plausibility to the model, they therefore suggested to add background knowledge to the memory system. This extended version of the model was explored in Nosofsky and Palmeri (1997a) as well as in Nosofsky and Alfonso-Reese (1999). Formally, the Background Noise Element version of EBRW (EBRW-BNE) is identical to the basic version of the model with one supplementary assumption. Before the simulated category learning begins, the exemplar memory of EBRW-BNE is assumed to contain a given number of existing instances that may enter the race and trigger a response. As such, they are “noise” in the system, because they are not systematically associated with the categories being tested. Thus, if any of these background noise instances bears a fortuitous resemblance to the item being categorized, it has the potential to drive the accumulator toward one of the boundaries, especially early in the learning phase. Thus, this background noise has, in theory, the potential to generate $c$ parameter values that differ from 1. Unfortunately, this does not turn out to be the case as we will now show.
The exemplar memory of EBRW-BNE is composed of the background noise elements
and one copy of each training item at block 1. We use the subscript G to denote the former.

\[ J_1 = \{ E_1, \cdots, E_i, E_{G1}, \cdots, E_{Gm} \} \quad (18) \]

Although the exact number of these background noise elements \( m \) is unknown, only their
summed similarity to a test item \( i \) is important to determine their impact on the steps taken by the
accumulator. From equations 7 and 18, we can determine that the expected time for a step
depends on the summed similarity of the test item \( i \) to all memorized category exemplars, and to
the set of background elements, so that at block \( N \), we have:

\[ E(T_s | i) = \frac{1}{N \left( S_{iA1} + S_{iB1} \right) + S_{iG} + \alpha} \quad (19) \]

The background noise elements are fixed before learning and so the similarity of item \( i \) to those,
noted \( S_{iG} \), is a constant independent of \( N \). Similarly, the probability of a step in direction of
boundary A given item \( i \), \( p_a \), depends on the summed similarity of \( i \) to those exemplars in
memory associated to A (all the items in category A and presumably half of the background
items), so that

\[ p_i = \frac{N \times S_{iA1} + \rho S_{iG}}{N \left( S_{iA1} + S_{iB1} \right) + S_{iG}} \quad (20) \]

in which \( \rho \) is the proportion of background elements associated with A (.5 was used in the
following figures as in Nosofsky and Alfonso-Reese, 1999). From this line of reasoning, \( a_i \) and
therefore \( E(N | i) \) can be derived.

Using a log-log plot, we can examine \( E(T | i) = E(T_s | i) \times E(N | i) \) (the scaling
parameters can be omitted without loss of generality). Figure 5(a) shows the predicted learning
curve along with a power curve having a curvature of 1. In log-log coordinates, a power function
EBRW and Learning curves

with zero asymptote becomes a linear function with a slope of \(-c\). In Figure 5(b), we show

\[ E(T \mid i) \] and a power function in log coordinates. As seen, \( E(T \mid i) \) is not linear, which indicates

that this curve is not a power function. Nevertheless, it tends rapidly toward linearity as the

number of training blocks increases. Moreover, the slope of the whole curve is always below 1

and the slope of the linear segment, exactly 1.\(^6\) Figure 5 highlights the role of the number of

blocks: if only two data points are available (for the specific parameters of that figure), the

observed slope is 0.8 and therefore, the beginning of the curve in normal coordinates would be

similar to a power function with an exponent of 0.8. However, whenever more than 10 blocks are

available, the slope tends to 1 and the learning curve is undistinguishable from a power function

with a curvature of 1. For large \( N \)s, the slope becomes exactly 1, since, \( N \rightarrow \infty \), Eq. 20 reduces

to the regular EBRW equation. In Figure 5(a), the power function most similar to the learning

curve of EBRW-BNE has a curvature of 0.92.

We explored under which conditions the curvature is much smaller than 1 by varying the

size of \( S_{iA} \) and \( S_{iB} \), relative to \( S_{iG} \). We examined three scenarios, manipulating the position of

the category boundaries: \( A = B \), \( A \gg B \) and \( A \ll B \), but they gave qualitatively identical results,

as described next. In order to maximize the chance of observing a curvature different from 1, we

used a small number of simulated blocks, namely \( N = 1 \) to 30 (compared to 150 in Nosofsky &

Palmeri, 1997a). The curvature was estimated as in the previous simulations.

Figure 6 shows the estimated curvature over 30 blocks as a function of the summed

similarities of item \( i \) to category A and the summed similarity of item \( i \) to category B for the

three scenarios examined. The summed similarities are relative to \( S_{iG} \), set as a baseline to 1. As

can be seen, to obtain a learning curve with a curvature smaller than 0.8, both \( S_{iA} \) and \( S_{iB} \), must
be considerably smaller than $S_{iG}$. In fact, with $A = B = 5$, $S_{iA}$ and $S_{iB}$ must both be $10^4$ times smaller than $S_{iG}$ to obtain a curvature of 0.2 (a curvature similar to the one obtained by the second participant of Nosofsky and Palmeri, 1997a). In other words, exemplar memory must contain background items that are thousands of times more similar to the test item $i$ than the training exemplars. This is of course unlikely and, in addition, has implications for training accuracies: the proportion of correct categorizations would remain close to random for hundreds of blocks.

The previous examination of EBRW-BNE shows that parameter $S_{iG}$ is identifiable relative to $S_{iA}$ and $S_{iB}$. It also shows the diagnostic value of learning curves to support or challenge models.

4- Possible EBRW enhancements and their relation to curvature

The EBRW model has already been modified in many ways. For instance, Palmeri (1997a) added a supplemental algorithm that served as a competitor to EBRW, to produce a race between a counting algorithm (see Logan, 1988) and the random-walk process in a numerosity judgment task. Because the algorithm’s effect can be seen as truncating the slowest RTs, its impact is to lower the upper part of the curve. Hence, when measured with the appropriate time scale, the observed curvature could be diminished, but never increased. This result would be compatible with the three participants of Nosofsky and Palmeri (1997a).

There is a straightforward modification that can be made to allow EBRW to predict various curvature parameters. Following Nosofsky and Palmeri’s (1997a) suggestion, the assumption that step times are sampled from an exponential distribution could be replaced by an assumption that they are sampled from a Weibull distribution (with the shape parameter free to
vary, see Cousineau et al., 2002), thus enabling the model to predict many different curvature
parameters.

The Weibull distribution is justified for two reasons. First, the Weibull is a generalization
of the Exponential distribution and includes that distribution as a special case. Second, the
Weibull distribution is the asymptotic distribution of extrema (Luce, 1986), as proven by the
Extreme Limit Theorem (Cousineau, Goodman and Shiffrin, 2002). Hence, if the process of
retrieving one exemplar constitutes a race (among, say, neural pathways), then that process
shows a Weibull distribution of retrieval times.

The basic version of EBRW assumes that each exemplar races to drive the next step in
the random walk. Nosofsky and Palmeri (1997a) originally assumed that the time to retrieve
exemplar $j$ when presented with test item $i$ (and thus perform one step in the random walk, $T_{R,ij}$),
followed an exponential distribution with a rate parameter $s_j$ (the computed similarity between
items $i$ and $j$). Hence, the exponential cumulative distribution function is equal to

$$Pr(T_{R,ij} < t) = 1 - e^{-s_j t},$$

and the winner of the race, assuming independence, is retrieved at
time $t$ with probability:

$$Pr(T_s < t \mid i) = Pr(\text{Min}[T_{R,ij}] < t \mid i)$$
$$= 1 - Pr(T_{R,ij} > t, \forall j \mid i)$$
$$= 1 - Pr(T_{R,i1} > t \mid i) \times \ldots \times Pr(T_{R,in} > t \mid i)$$
$$= 1 - \prod_{j \in J_i} 1 - Pr(T_{R,ij} < t \mid i)$$
$$= 1 - \prod_{j \in J_i} e^{-s_j t}$$
$$= 1 - \left( \prod_{j \in J_i} e^{-s_j t} \right)^N$$
$$= 1 - e^{-(N \sum s_j)t}$$
EBRW and Learning curves

1. (Feller, 1957). This is the result reported by Nosofsky and Palmeri (1997a). Time steps are also exponentially distributed with a parameter equal to the sum of the individual parameters.

2. It is known that the mean of an exponential distribution and its standard deviation are equal to the reciprocal of the distribution’s parameter $E(T_s | i) = 1 / (N \sum s_{ij})$. Consequently, the learning curves of the mean RT and that of the standard deviation in RT will have the exact same parameter value for curvature parameter $c$. This prediction should be familiar since Logan (1988, 1992) derived it when he developed his Instance-Based Theory of Automaticity. Nevertheless, this has never been indicated to apply to EBRW as well.

3. Following a suggestion made by Nosofsky and Palmeri (1997a, p. 270) and Logan (1988; 1992), a new version of EBRW can be devised in which the races to reach the categorical boundaries are assumed to be distributed according to the Weibull distribution rather than the exponential distribution. In $EBRW_{\text{weibull}}$, the cumulative distribution function is equal to

4. $\Pr(T_{R,ij} < t) = 1 - e^{-(s_{ij} t)^\gamma}$ in which $\gamma$ is a shape parameter altering the asymmetry of the distribution. We assumed a unique shape parameter because it is probably related to the physical implementation of the memory system (Tuerlinckx, 2004). If all exemplars are independent and follow a Weibull distribution with parameters $s_{ij}$ and $\gamma$, the fastest exemplar of the race will then be retrieved at time $t$ with probability:
EBRW and Learning curves

\[
\Pr(T_s < t | i) = \Pr(\text{Min}[T_{R,ij}] < t | i) \\
= 1 - \Pr(T_{R,ij} > t, \forall j | i) \\
= 1 - \Pr(T_{R,i1} > t | i) \times \ldots \times \Pr(T_{R,ijN} > t | i) \\
= 1 - \prod_{j \in J_N} 1 - \Pr(T_{R,ij} < t | i)
\]

\[
= 1 - \prod_{j \in J_N} e^{-(s_{ij})^\gamma} = 1 - \left( \prod_{j \in J_i} e^{-s_{ij}^\gamma} \right)^N \\
= 1 - e^{- (N \sum s_{ij}^\gamma)^\gamma} \\
= 1 - e^{- \left( \frac{N^{1/\gamma} \left( \sum s_{ij}^\gamma \right)^{1/\gamma}}{t} \right) \gamma}
\]

Hence, each time step also follows a Weibull distribution with parameters \(N^{1/\gamma} \left( \sum s_{ij}^\gamma \right)^{1/\gamma}\), and \(\gamma\).

As a consequence, the mean and standard deviation in RT are given by

\[
E(T_s | i) = \frac{1}{N^{1/\gamma} \sqrt{\sum s_{ij}^\gamma}} \Gamma \left( 1 + 1/\gamma \right)
\]

\[
SD(T_s | i) = \frac{1}{N^{1/\gamma} \sqrt{\sum s_{ij}^\gamma}} \sqrt{\Gamma \left( 1 + 2/\gamma \right) - \Gamma \left( 1 + 1/\gamma \right)^2}
\]

so that using the notation of Eq. 12:

\[
E(T_s | i) = k_2 N^{-1/\gamma} + \alpha
\]

in which \(k_2 = \frac{1}{\sqrt{\sum s_{ij}^\gamma}} \Gamma \left( 1 + 1/\gamma \right)\). Thus, the predicted curvature is \(c = 1/\gamma\) for both the mean and standard deviation. On the other hand, \(E(N | i)\) (Eq. 11) remains unchanged.

Two things are worth mentioning. First, response time distributions, when fit using a Weibull distribution, are typically best-fit using a parameter \(\gamma \approx 1.75\) (see, e.g., Cousineau, Engmann & Mestari, in press). If this measure is indicative of the underlying race process assumed in EBRW- BNE, the curvature of RTs should then be \(1/1.75\) or 0.57, a value closer to the mean curvature in the replication reported above (where it was 0.27). Second, a similar
parameter $\gamma$ was envisioned by Nosofsky (1986), where it was called a sensitivity parameter.

EBRW-BNE would then provide an alternative justification of this parameter.

Again, in this extension, both mean and standard deviation ought to have the same curvature. However, here $c$ is a function of the free parameter $\gamma$, and its value is therefore not restricted to 1. This last prediction regarding equal curvature of mean and standard deviation is interesting and can be further explored. Indeed, only parallel models can make this kind of prediction for *standard deviations*. Serial models would have equivalent predictions based on variance (Townsend and Ashby, 1983).

In the data from the replication, one participant had no clear curvature and this was also true when we examined his or her standard deviation. The second participant with the least curvature in mean RT (a curvature of 0.08; see Figure 3) had likewise no clear curvature when the standard deviations were examined. For the remaining ten participants, the correlation between the $c$s obtained on the mean RT and the $c$s obtained on the standard deviation of RT was 0.89. Hence, about 80% of the fluctuation in mean RTs across training also change the spread of response times, as predicted by EBRW$_{\text{Weibull}}$.

**5- Discussion**

**EBRW and learning**

We have discussed one problematic prediction from the first version of the EBRW model (the fixed learning rate parameter) and underlined the importance of time scale selection for fitting purposes. Those limitations were identified using only the learning curves. Hence, learning curves are quite informative and allow more predictions to be tested about models of learning than the traditional "end of training stage" statistics. In EBRW's case, as further
exposures to the stimulus set occur, the memory stores additional copies of the items. It was
shown that the net result on the summed similarities (the critical quantity in GCM and
consequently, in EBRW) increases linearly with the number of blocks. However, the
categorization responses are not based on the summed similarities, but on a ratio of summed
similarities. As such, the effect of additional exposure to the entire stimulus set is null. This
prediction made by EBRW is counter-intuitive. Thus, the learning curve for accuracy is probably
as important as the learning curve of the response time data when examining a model's fit.

**Novel predictions from the model**

Although many modifications could be suggested to upgrade EBRW, one of the great
merits of the model is that it placed the focus on learning curves. Indeed, Nosofsky and Palmeri's
(1997a, 1997b) model makes strong predictions regarding learning curves that can be tested
empirically. In turn, the generated results would place stringent constraints on future models of
speeded classification. Here are three such strong predictions:

1. Both EBRW and the Weibull-distribution EBRW models make the general prediction
   that both the mean and the standard deviation should have equal curvatures
   (Wagenmakers & Brown, 2007). This prediction was also part of Logan’s Instance-Based
   Theory of Automaticity and was shown to hold in the tasks studied in Logan (1988,
   1992). So far, the adequacy of this prediction has not been verified systematically in
categorization tasks. If it were shown to be general, this prediction would have strong
implications and would reject large classes of models that do not generate it.

2. Another strong prediction made by EBRW is that the learning curvature for all items
   should be identical. In other words, all items should be learned at the same rate. This is a
strong prediction as intuition suggests that there should be variability in the items’ learning rates. For instance, items that have a high degree of within-category similarity and a low degree of between-category similarity might exhibit faster learning rates than other items. Hence, the learning curvature might depend on which test item was presented (e.g., the curvature parameter for item i could be proportional to its summed similarity to other members of the category). Finding identical curvatures for all presented items would be an important finding, foreseen only by EBRW. Nosofsky and Alfonso-Reese (1999) did look at individual items' learning curves but focused on the commonalities between the curves rather than on their differences.

Similarly, the learning exponent does not depend on the composition of the category structure. Hence, categories whose exemplars are all clustered in one part of the psychological space, for example, should be learned at the same rate as categories that are composed of exceptions only. This prediction concerns only the learning rate; the asymptote and amplitude (parameters a and b from Eq. 1) may differ between these conditions. However, EBRW predicts that the rate at which the categories are acquired should be identical, no matter the difference in category structure. Again, this is a prediction that strongly contradicts intuition: finding empirical support for it would strongly support EBRW. Palmeri (1997a) contrasted two such conditions, reporting different mean curvatures in a numerosity judgment task; the question whether this result would hold in a categorical task remains open.
Conclusion

This whole discussion points to the importance of using learning curves to perform empirical tests of models. Psychological models of learning like EBRW often carry underlying assumptions that can be highlighted by the study of learning curves. For instance, EBRW predicts that item learning rates are equal both within and between categories. Also, it implies that categorization probabilities should be constant across blocks. Here, some of the predictions made by EBRW were explored. Similar analyses could be performed on many other psychological models, and large classes could be supported or rejected, thus accelerating the progression of psychological modeling research.
References


EBRW and Learning curves


Footnote

1 The demonstration is dependent on the test stimulus $i$ presented. However, the average of
several power functions with curvature $c$ is also a power function with curvature $c$ (Cousineau et
al., 2003; Estes, 1956). The demonstration therefore generalizes across all the items. In the
following, there are a number of possible confusions between the variables: $N$ is the block
number but also denotes (in bold) the number of steps to reach a boundary. There are also the A
and B category labels, not to be confused with the boundary positions $A$ and $-B$. We preserved
the original notation, as the context is generally not ambiguous.

2 For one participant, shown in the top-left panel of Figure 3 (slow learners), no learning curve
was visible (or learning was finished by the end of block 1). The average curvature is $0.301 \pm
0.06$ if we exclude this participant.

3 Because EBRW is a stochastic process model, one must expect variability in the model’s
predictions due to local minima during fitting and due to the random stimulus presentation order.
The key here is to show that the average curvature obtained when simulating the model is equal
to 1 (with a low standard error), and that the model and the procedure to find the best-fitting
parameters do not introduce any systematic biases.

4 Regarding the exponential curve, if the time scale is compressed by a factor of, say 5 as in
Nosofsky and Palmeri, 1997a, the curvature parameter will be five times larger and the other
parameters (asymptote and amplitude) will be identical. Nothing of the sort occurs for a scale-
dependent learning function (see Cousineau et al., 2003).
In all cases, the estimated amplitude of the curves (the parameter $b$ in Eq. 1) is larger because, had the performance on the very first block been available, it would presumably have been larger than the mean across the first 5 blocks.

To verify this, we computed the negative of the derivative for the log-log transformed equation $E(T | i)$. We found:

$$\frac{\partial \log(E(T|i))}{\partial N} = 1 + (S_{iB} - S_{iA}) e^{\frac{4A}{4} \frac{S_{iG}}{S_{iB} + S_{iG}}} \left( -B^2 (r^A - 1)^2 r^B - 2AB (r^A - 1)^2 r^B + A^2 (r^A - 1) (r^{A+B} - 1) \right) r^{A-1}$$

in which we use the following notation: $r = \frac{e^N S_{iA} + p S_{iG}}{e^N S_{iB} + (1-p) S_{iG}}$ and $q(x) = \frac{r^x + 1}{r^x - 1}$. 
Figure 1: The Training RTs from Experiment 1 of Nosofsky and Palmer 1997a, one row per participant. The dots are the mean RT; the full line is the best-fitting learning curve. In the left column, the learning curve is based on 30 training blocks; in the right column, the best-fitting learning curves are based on 150 blocks.
Figure 2. (top) Evolution of percent accuracy as a function of block of training; (bottom) distribution of the learning curve exponents obtained in the replication of Shin and Nosofsky (1992, Experiment 1). The next figure shows the individual RT curvatures.

Note. Mean curvature = 0.27 ± 0.06 (s.e.m.).
Figure 3. The learning curves for twelve participants obtained in the replication of Nosofsky and Palmeri (1997a, Experiment 1). The participants are subdivided into three classes based on the learning rate, arbitrarily termed "Slow", "Medium" and "Fast" learners.
Figure 3 (continued).

**Fast Learners**

- **Curvature** = 0.294
- **Curvature** = 0.566
- **Curvature** = 0.581
- **Curvature** = 0.704
Figure 4. Distribution of the estimated curvature parameter for 10,000 simulations of EBRW.
Figure 5(a). This figure shows a comparison of EBRW-BNE and a power function with an exponent of 1. (b) The same curves in log-log coordinates. The parameters used in this figure are $A = B = 4$, $S_{1a} = 2$, $S_{1b} = 1$, $S_{2a} = 1$ and $\rho = 1/2$. (c)
Figure 6: Estimated curvature parameter for 10 blocks or RT, as predicted by the EBRW-BNE model under three scenarios and for different $S_{iA_1}$ and $S_{iB_1}$, where $S_{iG}$ is arbitrarily set to 1.

$A = B = 5$

$A = 10, B = 1$

$A = 1, B = 10$