Nationaal Lucht- en Ruimtevaartlaboratorium
National Aerospace Laboratory NLR

NLR-TP-2002-446

Forecasting the learning curve for the acquisition of complex skills from practice

J.J.M. Roessingh, A.M.L. Kappers, J.J. Koenderink

Nationaal Lucht- en Ruimtevaartlaboratorium
National Aerospace Laboratory NLR

NLR-TP-2002-446

Forecasting the learning curve for the acquisition of complex skills from practice

J.J.M. Roessingh*, A.M.L. Kappers**, J.J. Koenderink**

* National Aerospace Laboratory NLR, ** Helmholtz Institute, Utrecht

This report is based on a presentation held at the 33rd European Conference on Mathematical Psychology, Bremen, 23 August 2002.

Customer:	National Aerospace Laboratory NLR
Working Plan number:	V.1.C.2
Owner:	National Aerospace Laboratory NLR
Division:	Flight
Distribution:	Unlimited
Classification title:	Unclassified
	August 2002

Summary

Three groups of novice trainees received extensive practice training on computer-based vehicle control tasks. The first group practiced with a primary manual control task and the second and third group practiced with more complex versions of this task involving secondary tasks.
Individual skill level during training was measured on the basis of times that trainees needed to successfully complete subsequent trials on the task. A total of 36,000 trial times were measured. The goal was to investigate whether the required training time for an individual to meet a particular level of competence could be predicted by extrapolation of that individual's learning curve.
Extrapolation of the learning curve over a substantial period requires (1) a valid analytical model and (2) a model for the random variations in individual performance. These requirements are met by the 'linear rate model', which specifies a learning curve function with one free parameter ('dead time'). The model also specifies the probability distribution of the random variations in performance. On this basis the dead time parameter can be estimated. The resulting model is used to predict the times that 16 individuals would need to complete 500 and on average 1800 task trials. The prediction errors are on average 0.46 and 4.1 percent with standard deviations of 10.3 and 13.7 percent, respectively. We discuss the utility of the model.

Contents

(30 pages in total)

1 Introduction

In this research, we construct and extrapolate learning curves of trainees who practice complex tasks and do not receive further instructional guidance during training. The learning curve is a measure of performance ('skill level') against a measure of practice.
As the measure of practice we choose the number of practice trials on a specific task. We investigate 'speed-tasks', which allow skill level to be measured principally by the time that a trainee needs to complete a practice trial, once successful completion of the task can be taken for granted. For the time needed to complete each trial, we use the term 'trial time' throughout this study. We do not use the term 'response time', because a trial on a complex task may require multiple responses.
By extrapolating the individual learning curve function beyond the trial times on which its fit is based and testing the extrapolation against newly available trial time data, we obtain an appreciation of the underlying model, in this case the 'linear rate model' (Roessingh, Koenderink & Kappers, 2001a, 2001b). We may thus have a theoretical model that accurately describes the trend in trial times over a number of past trials. However, this does not guarantee that the model is able to predict the trend in a series of trials in the future.
The objective of the current research is to quantitatively predict future trial times $\{T_{n+1}, T_{n+2}, \ldots\}$ of individual trainees, and associated quantities, such as the practice time to achieve a criterion, on the basis of their previous trial times $\{T_1, .., T_n\}$.

Complex tasks used in this study

The experimental tasks are increasingly difficult versions of the Space Fortress (SF) game. SF is a Personal Computer game that was specifically designed for the study of complex skills acquisition. Experimental research with the SF game has been documented in a special volume of Acta Psychologica (edited by Donchin, Fabiani & Sanders, 1989) and in various later research publications, such as field studies at flight schools where SF has been used in flight training (e.g. Gopher, Weil & Bareket, 1994, Hart & Battiste, 1992, Vidulich, McCoy & Crabtree, 1995). All three versions of SF used in this study are sufficiently complex and interesting for trainees to guarantee a very long skill acquisition process. SF also allows for reliable measurement of valid performance measures, in our case thousands of subsequent trial times per trainee. We describe the task in more detail in the Method section.

Overview of the study

In the remainder of this paper we shortly describe the 'linear rate model', which gives an explanation for the acquisition of complex skills from practice. We derive a learning curve function from this model (in terms of subsequent trial times as a function of practice trials) and a probability distribution for subsequent trial times. We employ the latter distribution to estimate the free parameter (the dead time parameter T_d) of the learning curve function. We propose an iterative algorithm for the estimation of dead time. In the subsequent Method section we describe the experiment. We start the Results section with a summary of the data. Subsequently, we provide estimates for the dead time parameter for each individual, using the previously described method. Then, we present the fits of the extrapolated learning curve functions to the trial times of individual trainees in numerical and graphical form. We apply these extrapolations to predict the practice time needed to perform a criterion number of future trials. Finally we discuss to what extent individual skill acquisition can be predicted.

2 Theory

A learning curve function with one free parameter

A functional expression for the learning curve seems a necessary condition for the purpose of predicting ‘future’ trial times of an individual. However, since trial times appear to vary randomly from one trial to the next, trial times for an individual do not follow a smooth analytical function with a limited number of parameters. Hence, the trial times have to be considered as stochastic variables and one can only predict an expected trend in average trial times. This trend is generally found to decrease with practice. An essential feature of a learning theory is that it should account for this decrease. To arrive at a functional expression for the learning curve, we developed the ‘linear rate model’ (Roessingh et al, 2001a, 2001b). This model considers learning of a complex task from practice as: (1) exposure to the task, (2) accumulation of information on how to attain the task goals, which accumulated information we call ‘perception-action patterns’, and (3) retrieval of perception-action patterns. These three operations are not considered distinct – each one leads to the others. Perception-action patterns must be considered as mere abstractions, to provide an adequate description level for learning phenomena.

The proposed model is based on preference for successful exposures to the task, such that patterns that represent these successful exposures will be accumulated. During each exposure to the task, a randomly retrieved pattern guides this exposure in terms of perceptions and actions required to finish the task. If the current exposure to the task is deemed more successful than the guiding pattern, a new pattern will be created. This new pattern represents the current exposure and may guide future exposures to the task. Alternatively, when the current exposure is unsuccessful relative to the guiding pattern (i.e. when the guiding pattern has better prospects than the actual outcome of the exposure), the guiding pattern will create an approximate copy of itself, which may also guide future exposures to the task. We refer to this hypothetical mechanism as ‘paired selection’.

Thus, a new pattern will generally be more successful than an old pattern. More specifically, with random retrieval of one pattern, and creation of one new pattern on each exposure, as the result of one paired selection, the new pattern will be on average twice as successful as the average pattern that guided previous exposures.

In a speed-based task, successful patterns lead to faster trial times. Thus, when we consider each complete trial on the task as one exposure to the task, paired comparison leads to trial times whose expected values are exactly half the average trial time of all preceding trials. In formula:

$$\hat{T}_{n+1} = \tfrac{1}{2}\langle T \rangle_n \,. \qquad (1)$$

$\langle T \rangle_n$ denotes the average trial time over *n* trials, and $\hat{T}_{n+1}$ denotes the expected trial time of the subsequent trial. At the first trial *n*=1, and $\langle T \rangle_1 = T_1$.

We refer to the function of equation 1 as a ‘progressive average function’. However, this model for the learning curve is not yet complete. Tasks require a certain amount of motor action from the human operator, which will give rise to signal transport delays, and the nature of the task may also involve delays in the task environment. Thus, if the number of trials *n* goes to infinity, the trial time will eventually asymptote at a constant non-zero level, which level is not sensitive to further learning. We refer to this level as ‘dead time’ T_d. If we generalize the progressive average function of equation 1 with a free parameter T_d to cater for constant dead time, it becomes:

$$\hat{T}_{n+1} - T_d = \tfrac{1}{2}\left(\langle T \rangle_n - T_d\right), \tag{2}$$

and hence,

$$\hat{T}_{n+1} = \tfrac{1}{2} T_d + \tfrac{1}{2}\langle T \rangle_n . \tag{3}$$

Equation 3 has an asymptote at T_d, because the average $\langle T \rangle_n$ of the expected trial times will approach T_d when *n* goes to infinity.
We previously investigated (Roessingh et al., 2001a) the correspondence between the progressive average function of equation 3 and the two-parameter power function:

$$\hat{T}_n = T_d + b \cdot n^{-\frac{1}{2}}, \tag{4}$$

The two functions are equivalent up to a Taylor correction in *n* for the first few trial times. However, the power law of equation 4 has an extra scale parameter *b*. For the purpose of extrapolating the individual learning curve, we prefer the progressive average function, because it has only one free parameter T_d and thus reduces problems related to parameter estimation.

<u>A model for random trial-to-trial errors</u>

The progressive average function of equation 3 has one free parameter T_d, which needs to be estimated for each trainee separately from the empirical series of trial times $\{T_1, ..., T_n\}$. However, methods for parameter estimation generally require information on the probability distribution of the random trial-to-trial errors. Thus, besides a functional expression for the decreasing trend, information concerning the distribution of the trial-to-trial errors is required for reliable estimation of the free parameter.
We analyzed the random trial-to-trial error during complex skill acquisition in two experiments (reported in Roessingh et al., 2001a, 2001b) and we found that random trial-to-trial error is multiplicative. Moreover, we established that both scale and shape of the probability distribution of the trial-to-trial error change during the learning process.
The instance theory of automaticity, put forward by Logan (1988, 1992), provides an explanation for the distribution of the random trial-to-trial error. This theory models a single

response time during a learning process as the result of a ‘race’ between many memory instances (memory traces of previous exposures to the task) and a mental algorithm, the latter being a relatively slow series of steps that takes the learner through the task. As more instances accumulate with practice trials, it becomes increasingly likely that one of these instances will win the race from the algorithm. Eventually, the time preceding a response is governed by the time needed to retrieve the ‘winning’ instance from memory, rather than by the execution of an algorithm. By formalizing the horse race as a stochastic process, Logan derives the distribution of subsequent response times (1992, p. 886, Eq. 4), which is a Weibull distribution. More specifically, the instance theory predicts that the scale of the Weibull distribution decreases with a power function of trials. This power function has an exponent that is the reciprocal of the shape parameter.

We simplified Logan’s race mechanism to the previously described paired selection mechanism, which can be considered as a race mechanism with only two racers. Using Logan’s line of reasoning, a race mechanism with only two racers predicts that the trial-to-trial errors (and hence the trial times) that are lower than the expected value follow a Weibull distribution with shape parameter 2. A precise prediction of trial times that fall above the expected value could not be made. By introducing a random error variable ε_n, with expected (mean) value 1 and minimum value 0, the progressive average function of equation 3 can be rewritten as:

$$T_{n+1} = T_d + \tfrac{1}{2}\left(\langle T \rangle_n - T_d\right)\varepsilon_n, \qquad (5)$$

We assume that the random error variable ε_n has the previously described probability distribution, which is the Weibull distribution with shape parameter 2 at the low tail and a unspecified distribution at the high tail. The lack of specification of the overall distribution restricts the statistical methods with which we can estimate the free parameter T_d.

Estimating the dead time parameter

In order to predict the trend in a series of stochastic trial times $\{T_1, .., T_n\}$ on the basis of the model of equation 5 we need to estimate the dead time parameter T_d, which eventually causes the learning curve to saturate at a trial time that is larger than zero. Reliable estimation of the ‘asymptotic trial time’ (when practice time goes to infinity) from a series of trial times is a long-standing problem in the research of learning (which will be re-addressed in the discussion section). Common techniques, such as those based on fitting a learning function with least-squares regression fail to produce reasonable estimates. A frequently encountered problem is that these general techniques yield estimates that are largely biased, resulting in physically implausible (e.g. negative) values for dead time. A primary reason might be that the minimum condition for application of these techniques, a stable and recognized distribution of residual error, is not satisfied.

As a possible solution to this problem we propose an iterative method for estimating the dead time parameter T_d from a series of trial times. We fit the Weibull distribution with shape parameter 2 to the smallest value of the empirical distribution of the random error variables ε (see equation 5). The assumption that smallest value of the empirical distribution has the specified Weibull distribution is based on the paired selection mechanism, which is assumed in the linear rate model. Rather than fitting the Weibull density function *f* or the cumulative distribution function *F*, we fit the hazard rate *h*:

$$h(\varepsilon) = \frac{f(\varepsilon)}{1 - F(\varepsilon)}. \quad (6)$$

Roughly speaking, the hazard rate of the random error variable ε can be interpreted as the probability of immediate completion of the trial, given that the increasing error has a value of ε already. The random error ε for each trial can be approximated by using the progressive average function of equation 5, giving:

$$\hat{\varepsilon}_n = \frac{2(T_n - T_d)}{\langle T \rangle_n - T_d}. \quad (7)$$

The method starts by determining the minimum trial time $T(1)$ from the series $\{T_1, T_2, \ldots, T_n\}$ as a first estimate $\hat{T}_d$. We denote the trial number at which $T(1)$ occurs by $n(1)$. With the first estimate for T_d, equation 7 can be used to calculate a series of n approximate random errors. Let $\hat{\varepsilon}(i)$, $i = 1, \ldots, n$, be the rank ordering of errors $\hat{\varepsilon}$, from the smallest error $\hat{\varepsilon}(1)$ to the largest error $\hat{\varepsilon}(n)$ and with $\hat{\varepsilon}(i)$ the error having the i^{th} rank. We approximate the hazard rate $h(\varepsilon(i))$ by the empirical hazard rate $\hat{h}(i)$ for the i^{th} interval $\Delta\hat{\varepsilon}(i) = \hat{\varepsilon}(i+1) - \hat{\varepsilon}(i)$, which can be calculated by (Singpurwalla & Wong, 1983):

$$\hat{h}(\varepsilon(i)) \approx \hat{h}(i) = \frac{i}{\sum_{j=1}^{i} (n - j + 1) \cdot \Delta\hat{\varepsilon}(j)}. \quad (8)$$

Low values of $\varepsilon(i)$ are assumed to have a Weibull distribution with mean 1 and shape parameter 2. The theoretical hazard rate of $\varepsilon(i)$ is then given by the Weibull hazard function:

$$h(\varepsilon(i)) = \tfrac{1}{2}\pi\, \varepsilon(i). \quad (9)$$

By setting the empirical hazard rate of equation 8 equal to the theoretical hazard rate of equation 9, new (fitted) values for the errors $\varepsilon(i)$ can be calculated, in particular the smallest error $\varepsilon(1)$:

$$\varepsilon(1) = \frac{2}{\pi \cdot n \cdot \Delta\hat{\varepsilon}(1)}. \tag{10}$$

Now, a new $\hat{T}_d$ can be calculated by again using the progressive average function of equation 5:

$$\hat{T}_d = \frac{2 \cdot T(1) - \langle T \rangle_{n(1)} \cdot \varepsilon(1)}{2 - \varepsilon(1)}, \tag{11}$$

in which $\langle T \rangle_{n(1)}$ is the mean of trial times T_i, $i = 1, \ldots, n(1)$. Thus, once we have determined the empirical hazard rate $\hat{h}(1)$ we can obtain a new estimate $\hat{T}_d$. Equation 7 can then be re-applied to calculate a new series of approximate random errors $\hat{\varepsilon}(i)$. The method can be repeated until the difference between subsequent estimates $\hat{T}_d$ vanishes. In practice, a few iterations are sufficient and the method is generally convergent. To obtain a more reliable $\hat{T}_d$ than one that would be solely based on $\hat{h}(1)$, we calculate $\hat{h}(1)$ to $\hat{h}(5)$ by means of equation 8 and use the median value to estimate $h(\varepsilon(1))$.

3 Method

Tasks

In all three different versions of the SF-game used in this research and described below, the display contains a rotating fortress in the center and a maneuverable spaceship, which has a starting position in the lower right corner of the display. The trainee controls the spaceship's flight with a joystick. The trajectory of flight can be controlled by rotating the ship and applying thrust (which causes the ship to accelerate). The ship continues to fly in the direction in which it is pointing, unless it is rotated and thrust is applied. This 'control law' significantly contributes to the complexity of the task, as it is neither intuitive nor easy to learn for novice trainees.

Task 1 is a subset of the SF-game (Mane & Donchin, 1989). This specific task was used previously by Frederiksen & White (1989). The trainee controls the spaceship's flight with a joystick and fires missiles from the ship by pressing a fire button on top of the joystick. The trainee's task is to attack the fortress by hitting it ten times with a missile, at intervals of at least 250 ms, before destroying it with a burst of two shots (fired at an interval of less than 250 ms).
The fortress defends itself against the ship. It does this by rotating to face the ship and then trailing the ship's movements while firing shells at it. When the ship is hit for the fourth time by a shell from the fortress, it is returned to its starting position. When this happens, the shot counter, which counts the hits scored against the fortress, is set to zero. A trial on the task finishes as soon as the fortress is destroyed.

Task 2 is a more complex version of Task 1. Additionally, the fortress is protected by moving 'mines' which emerge on the display periodically. These mines chase the ship. Unless the trainee takes action, these mines will hit the ship. Moreover, when a mine is present on the display, missiles fired at the fortress have no effect. Thus, the mine has to be eliminated by a missile immediately. However, if the trainee fails to hit the mine within 10 seconds, the mine disappears from the screen automatically. The interval between the disappearance of one mine and the appearance of the next is four seconds, during which time the trainee can fire at the fortress. When the ship is hit for the fourth time by either a mine or a shell from the fortress, the ship is returned to its starting position and its shot counter is set to zero. As with Task 1, a trial on the task finishes as soon as the fortress is destroyed.

Task 3 is the full SF-game. In addition to Task 2, the trainee has to distinguish between two types of mines, and react accordingly. The more difficult mine can be identified by a letter that appears in the information panel at the bottom of the screen (prior to each five minute block of play, the trainee is presented with a new set of three letters that are used to identify 'difficult' mines). Appearance of a difficult mine requires the trainee to press the right ('identification') button on the mouse twice with an interval of 250-400 ms before the mine can be destroyed by a missile. The 'easy' mine, as in Task 2, can simply be destroyed by hitting it with a missile without pressing the identification button. However, falsely pressing the identification button for an easy mine makes this mine invulnerable for missiles, such that it cannot be eliminated and will either hit the ship or automatically disappear after 10 seconds. Since missiles fired at the fortress have no effect when a mine is present, the trainee can choose whether to avoid the invulnerable mine and wait for it to disappear or let it damage the ship. Another complication in this task is that the supply of missiles is limited, and the stock has to be monitored in the information panel at the bottom of the screen. Extra supply can be obtained by using 'resource opportunities'. The availability of these opportunities are indicated by a random sequence of symbols (&, #, $, %, !, etc.) which appear in the center of the display (beneath the fortress). When the $ symbol appears for the second time in a row, the trainee can get extra missiles by clicking the middle button of the mouse. As with the less complex tasks, a trial finishes as soon as the fortress is destroyed.

Trainees

Sixteen male university undergraduates aged between 20 and 23, with normal vision, participated in the study. Trainees were recruited via an advertisement in the University magazine of Utrecht University. In total 36 trainees were selected from a larger group of 51 candidates by means of the Aiming Screening Task, a task that is known to be a reasonable predictor for training success on this task (see Foss, Fabiani, Mane & Donchin, 1989). A minimum aiming screening score of 740 points was required to participate in the study. As the current study is part of a larger training study, the sixteen trainees in the current study are a balanced subset of the full set of 36 trainees who participated in the larger study. The subset has the same average screening score (870 points) as the full set, and each trainee with an above-average screening score is paired with a trainee with a below-average screening score. None of the trainees reported playing video games for more than 4 hours per week. Trainees were paid 30 euro per day plus a bonus of 68 euro upon completion of the experiment.

Procedure

We selected the conditions from the larger experiment in which trainees practiced with only one task (either Task 1, Task 2 or Task 3) and received no previous practice training on a different task. Eight trainees practiced with Task 1, two trainees practiced with task 2 and six trainees practiced with task 3. The different numbers of trainees in each of those three conditions are the result of the design of the larger study, but are of minor interest for the current study as the focus is on individual measures rather than group measures.
In the course of the experiment, each trainee received practice training on the task to which he had been assigned. The trainees completed on average 2300 trials (ranging from 1273 to 3857 trials). To this end, we scheduled eight training days over a five-week period for each trainee. During a training day, the trainee would complete three training sessions consisting of eight blocks of five minutes each, separated by two breaks of twenty minutes. The effective time-on-task was thus forty minutes per session and 120 minutes per day. Trainees were allowed to take

one-minute breaks between five-minute blocks. The data collected with Task 1 and Task 3 have been reported previously in Roessingh et al. (2001a, 2001b).

Software and equipment

The experiment room contained individual computer stations in separate cubicles. Each computer station was equipped with a PC and a joystick of type FlightStick (CH-products). The joysticks were modified so that they could be connected to an A/D converter card (DataTranslation) in the PCs. The fire-button on the joystick and the three other response buttons were connected to a timer card in the PC. A camera system was installed in the cubicles to control the course of the experiment.
The original SF software was made available by the Dept. of Psychology, University of Illinois at Urbana-Champaign. To facilitate Task 1 and Task 2, the software was modified to remove the specified components of the full SF-game. The software was also modified to record additional parameters, most importantly total time-on-task and trial-times, with a timing accuracy of 50 milliseconds.

Further training materials

After screening and well before the start of the experiment, the trainees received the instruction booklet for the SF game by mail at their home address. This instruction booklet specified the rules of the game and explained control of the space ship. No reference was made to specific tactics or strategies. The trainees were instructed to study the booklet carefully before the experiment began.

4 Results

Table 1

Grouped summary of trial time data over trials 1-500 and trials 501-1000[1]

Task	N[2]	Trials	Max. trial time[3] [s]		Min. trial time[4] [s]		Coefficient of Variation (CV)[5]		Practice time[6] [minutes]	
			M	SD[7]	M	SD	M	SD	M	SD
1	8	1-500	220	149	7.9	2.1	0.82	0.24	190	78
2	2	1-500	733	-	7.3	-	1.19	-	291	-
3	6	1-500	1048	320	9.2	2.8	1.38	0.33	407	87
1	8	501-1000	62	20	6.7	1.3	0.45	0.13	116	40
2	2	501-1000	64	-	6.5	-	0.46	-	152	-
3	6	501-1000	92	21	7.4	0.8	0.48	0.08	187	28

Brief description of trial times

Table 1 summarizes the trial time data, with separate figures for the first 500 trials (trial 1 to 500) and the second 500 trials (trial 501 to 1000). The trial time data have been summarized per task, task 1 being the least complex task, etc. In appendix F we provide the corresponding data for individual trainees. The mean M and standard deviation SD in table 1 represent group figures. For example, the first column in table 1 reports the maximum observed trial time. The maximum trial time of the eight trainees that trained with task 1 is on average 220 seconds during the first 500 trials with a standard deviation of 149 seconds. The standard deviations reveal that initially the variation in the maximum trial time between individuals with the same task is high. Maximum trial times are much higher with more complex tasks and strongly decrease in magnitude with practice trials. Conversely, the minimum observed trial times vary only slightly between individuals and hardly increase with more complex tasks. Moreover, with an increase in practice trials the minimum trial times drop only slightly.

[1] individual figures are reported in appendix F.
[2] number of trainees (16 in total) on each task/group.
[3] group mean (M) and group standard deviation (SD) of max. individual trial times.
[4] M and SD of minimum individual trial time.
[5] M and SD of ratio of standard deviation and mean (σ/μ) in 500 individual trial times.
[6] M and SD of time needed to complete 500 trials. Practice time is the sum of trial times.
[7] SDs are not reported for task 2 (N=2).

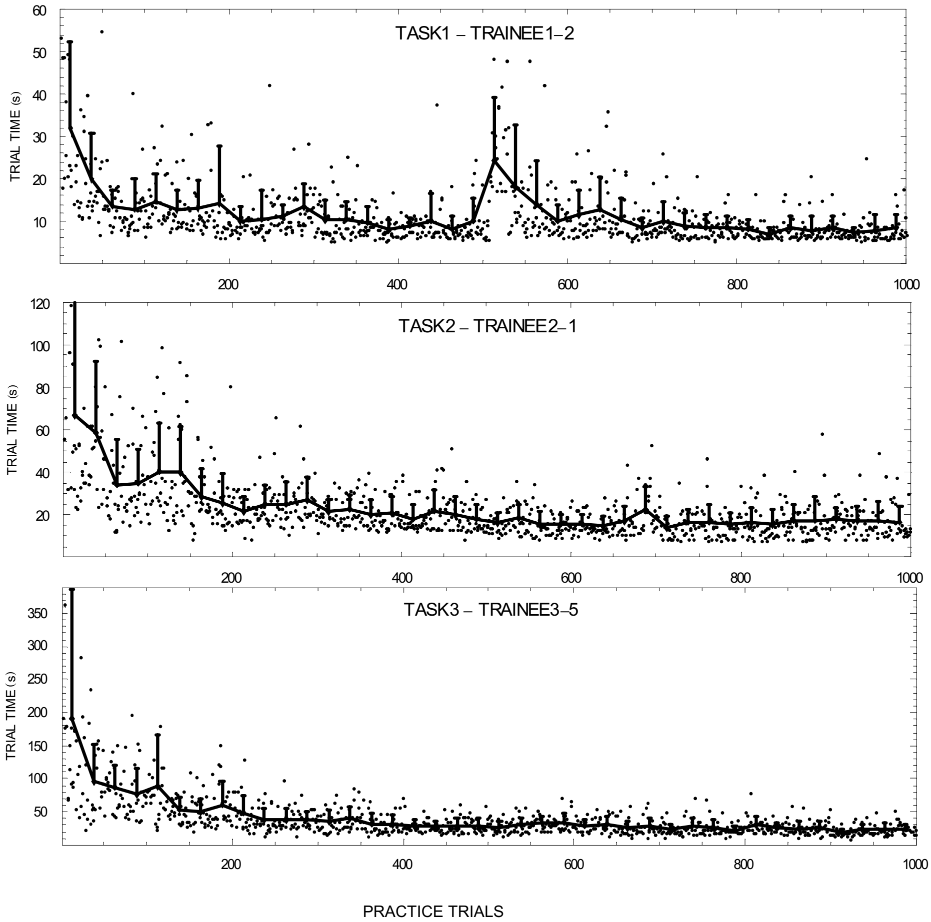

Figure 1: Trial times plotted against trial number for trainees from different task-groups. The line connects average trial times of subsequent series of 25 trials. The error bars denote the standard deviation in each series. Note the different time scales for the vertical axes.

The mean coefficient of variation (CV, the ratio between standard deviation and mean in 500 individual trial times) is initially higher for more complex tasks. However, for the second 500 trials, the CV is considerably lower for all three tasks and the dependence of the CV on task complexity has vanished. Roughly speaking, the CV for the first 500 trials is in the order of 1.0 or larger and drops to a value in the order of 0.5 in the second 500 trials. With more basic tasks, Logan (1992) reports considerable lower CV's, in the order of 0.2 to 0.4. For example, the CV for a dot counting task was 0.342. It must be noted that the CV relies on the standard deviation, which is based on squared deviations and thus emphasizes extreme deviations. Even if the data contain relative few extreme deviations, the CV will be determined predominantly by these deviations. Since with more complex tasks the maximum trial times are relatively large with respect to the mean, this may explain the high CV's in the first 500 trials of our data. Finally, the last column of table 1 reveals that the second 500 trials took approximately half the practice time needed for the first 500 trials, irrespective of task complexity.

In figure 1, the first 1000 trial times have been plotted for one trainee of each group, as representative examples. The trial times of all trainees show a decreasing trend and these trends seem to reach an asymptote. Figure 1 also reveals that trial times are more likely to be extremely positive than extremely negative with respect to an average value. Thus, trial times have a probability distribution that is skewed towards small values and with a long high tail. At first sight, there do not appear to be any specific features that distinguish the three plots from each other, apart from seemingly random deviations from the average trend and the differences between the trainees (tasks) in the time scale at which trials are completed.

Dead time estimates

The estimates for T_d, using the estimation method outlined in the theory section, are presented in table 2. We report estimates for each trainee based on the first 500, the first 1000 trials and all trials (on average 2300 trials per trainee). The 95 percent confidence intervals reveal which estimates are significantly ($p<0.025$) larger than zero.
If dead time estimates are based on 500 trials, only 25 percent of the estimates are significantly larger than zero. When based on 1000 trials, roughly half are significantly larger than zero. When based on all trials (on average 2300 per trainee), roughly 75 percent of all estimates are significantly larger than zero. In particular with the more complex tasks 2 and 3, the confidence intervals reflect large uncertainties. Thus, the method requires many observations to reliably determine dead time. Therefore, the differences between individuals with respect to dead time T_d are seldom significant. Only for Trainee 1-2 we estimated a dead time (3.94 s) that was significantly smaller ($p<0.025$) than that of Trainee 1-1 (5.13 s) and Trainee 1-4 (5.84 s).

The rules of task 1 (which is also the basis for task 2 and task 3) determine that each trial contains at least 2.5 seconds of waiting time (see the task description) and this waiting time biases T_d. Potentially, Trainee 1-2, whose estimated T_d is 3.94 s, is able to execute Task 1 in the intervals between shots (ten times 250 milliseconds) and a net 1.44 seconds. A trial time of 4.45 s has actually been observed in the data of Trainee 1-2.
As indicated by the last column of Table 2, even task 3 can be completed within 7.5 seconds by all trainees that practiced with this task. Given the complexity of the task, which includes actions like ‘applying thrust to accelerate/decelerate the space ship’, ‘aiming’, ‘firing with constant intervals’, and ‘monitoring the shot counter’, such short trial times are quite astonishing.
All estimates of T_d in table 2 are higher than 2.5 seconds and always lower than the lowest value actually observed in an individual series. Clearly, the method for estimating T_d yields plausible values.

Table 2

Estimates of dead time parameter T_d for each trainee.

	Basis = 500 trials		Basis = 1000 trials		Basis = all trials	
Task-Trainee	T_d[8] (estimated)	95% c.i. in T_d[9]	T_d[10] (estimated)	95% c.i. in T_d	T_d[11] (estimated)	95% c.i. in T_d
1-1	5.94	2.84 – 6.30	5.32	2.73 – 5.56	5.13	4.50 – 5.14
1-2	4.75	2.00 – 5.20	4.87	2.86 – 5.15	3.94	3.33 – 4.33
1-3	5.03	-3.15 – 5.81	5.49	2.88 – 5.54	4.67	3.30 – 4.68
1-4	5.03	0.98 – 9.10	6.86	3.52 – 6.94	5.84	4.63 – 5.88
1-5	6.39	0.85 – 6.78	5.54	1.82 – 6.28	5.37	3.83 – 5.56
1-6	6.57	-3.63 – 8.73	5.44	-0.44 – 7.52	6.16	3.48 – 6.50
1-7	11.01	-5.38 – 11.62	8.50	-0.02 – 9.20	7.10	3.14 – 7.29
1-8	7.95	-14.46 – 8.69	6.43	-9.35 – 6.82	6.11	-0.53 – 6.37
Mean	**6.58**		**6.06**		**5.54**	
2-1	7.24	-10.62 – 7.95	6.22	2.49 – 7.93	7.81	3.45 – 7.99
2-2	4.05	-14.57 – 5.71	4.93	-9.71 – 5.62	5.17	-9.85 – 5.67
Mean	**5.65**		**5.58**		**6.49**	
3-1	7.54	2.11 – 8.53	7.18	-3.81 – 7.22	5.62	4.26 – 5.71
3-2	2.92	-16.75 – 4.17	2.55	-13.98 – 4.57	4.03	-12.83 – 4.46
3-3	6.25	-3.38 – 7.97	6.83	0.378 – 7.04	5.77	2.11 – 5.95
3-4	5.44	-4.04 – 8.19	5.90	-2.35 – 6.48	5.63	1.02 – 5.72
3-5	5.82	-17.14 – 12.74	6.21	-1.52 – 7.64	7.23	-0.03 – 7.51
3-6	6.99	-29.38 – 10.22	6.53	-2.88 – 8.20	6.62	-0.50 – 6.96
Mean	**5.83**		**5.87**		**5.82**	

Extrapolation of the learning curve

The progressive average function (equation 3) predicts trial times from preceding trial times. We now test this model against individual trial time data. To standardize the analysis for the three different tasks and groups, we estimate dead time parameter T_d on the basis of the first 500 trials of each trainee. We estimate each subsequent trial time by recursive use of equation 3, without feeding new observations into the model. We estimate all remaining trial times (on average 1800) of each trainee in this fashion. In figure 2, we represent these results for three trainees graphically. To plot trial times T_n against trial number n, we used both linear and

[8] Estimates of T_d and 95 percent confidence intervals (c.i.) in these columns are based on the first 500 trials of each trainee.
[9] 95 percent confidence intervals. When the lower bound is larger than zero, T_d is significantly larger than zero ($p<0.025$).
[10] Estimates and intervals in these columns are based on the first 1000 trials.
[11] Estimates and intervals in these columns are based on all trials (on average 2300 trials per trainee).

logarithmic co-ordinates. Since the latter co-ordinates markedly straighten the curves, they provide a more balanced view of the deviations.
We evaluate the extrapolations by calculating the lack-of-fit to the observed trial times. Roughly speaking, there may be two reasons for the lack-of-fit of the extrapolation. First, the model is unable to predict 'unexpected' variations in the data. Second, the estimate of parameter T_d (table 2) may be inaccurate. In appendix B we report the following four lack-of-fit measures: (1) Mean Percentage Error, (2) Mean Absolute Percentage Error, (3) Root of Mean of Squared Errors and (4) Coefficient of Determination R^2. These will be explained in detail hereafter. These lack-of-fit measures have been calculated for the description of all trial times, and for two 'extrapolation periods': (1) trials 501-1000 and (2) all trials beyond trial number 500. In table 3, we report averaged lack-of-fit measures by means of a summary. Table 3a summarises lack-of-fit of the description, table 3b summarises lack-of-fit of 500 trials extrapolation and table 3c summarises lack-of-fit of all trials beyond trial number 500, i.e. 'the full extrapolation'.

<u>Mean Percentage Error (MPE)</u>. An obvious purpose of extrapolation is to evaluate the prospects of a trainee, for example in terms of the trend in trial times over some future period (the extrapolation period) or, equivalently, in terms of trials or training time required until the trend in the curve exceeds a certain criterion skill level. In both examples, the most appropriate lack-of-fit measure is the sum of residual errors in the extrapolated trial times (i.e. predicted trial time minus observed trial time, summed over all extrapolated trials). The sum of errors provides us with the gross error in predicted training time. To obtain a general reference, we compute the relative sum of errors by dividing the sum of errors by the sum of observed trial times. Multiplication by a factor hundred yields the Mean Percentage Error (MPE, StatSoft, 2001). These MPEs, averaged per task, are included in table 3.
The upper panels of figure 2, depict the extrapolation of the learning curve for Trainee 1-2, who practiced with task 1 (the least complex task). This extrapolation is the second worst of all extrapolations (see Appendix B), with an MPE of –23.7 percent over the extrapolated part of the curve.

As is apparent from figure 2, the lack-of-fit for Trainee 1-2 is caused primarily by the inability of the model to predict extremely large trial times. The peaks in the trial times of Trainee 1-2 possibly indicate temporary loss of skill of this trainee (e.g. due to forgetting during days on which no training took place).
Trainee 2-1, in the middle panels of figure 2, provides an example of a fairly accurate extrapolation, with an MPE in the full extrapolation of –1.28 percent.
The data for Trainee 3-5, who practiced the most complex task 3, is depicted in the lower panels of figure 2. The MPE is here 3.87 percent. Overall, the MPEs in the 500 trial extrapolations have a mean value of only -0.46 percent (standard deviation 10.3 percent). The MPEs in the full extrapolation (1800 trials on average) have a mean value of 4.07 percent (standard deviation 13.7 percent).

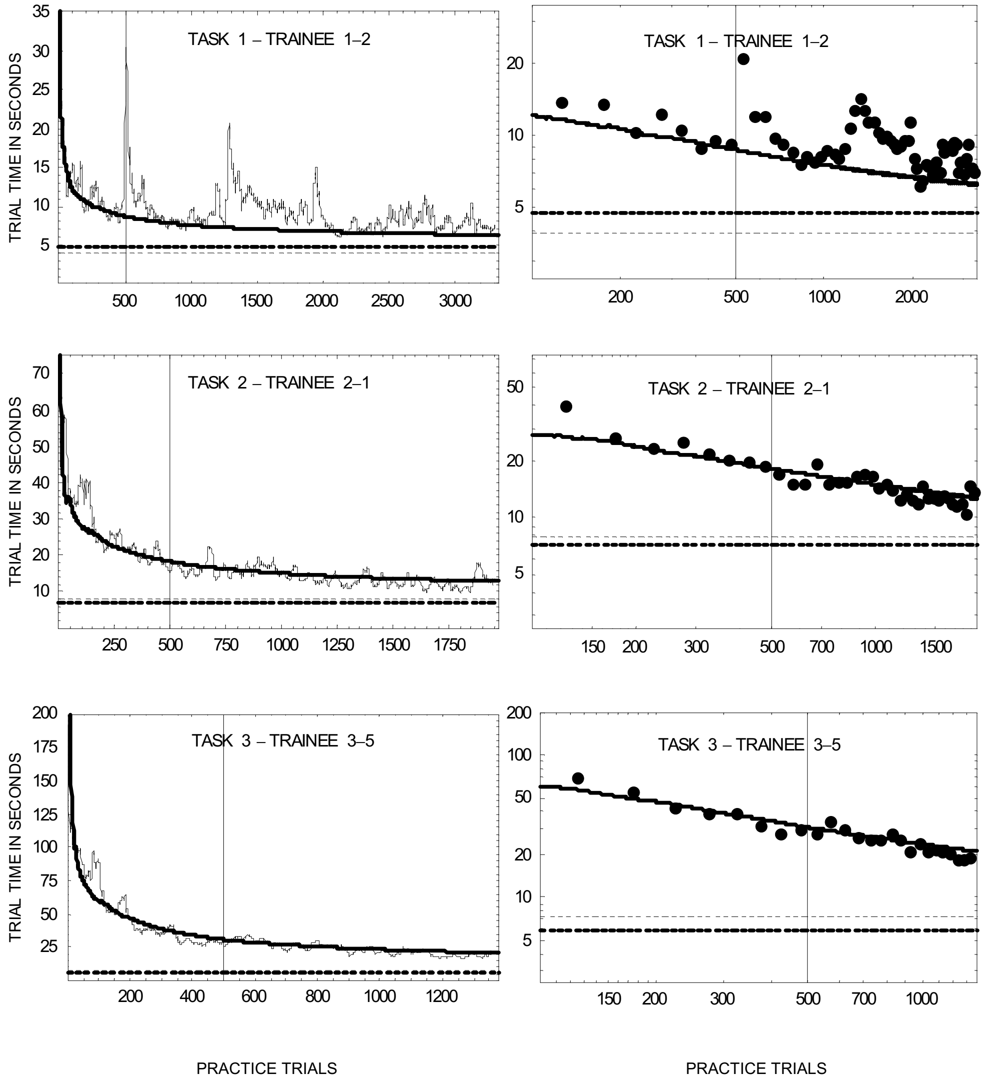

Figure 2: Extrapolations beyond 500 trials (vertical line) of the one-parameter model (equation 3) for three trainees (each with a different task) in linear and logarithmic coordinates. The jagged lines in the linear plots are the moving averages of observed trial times (window of 25 trials). The dots in the logarithmic plots are the average trial times for subsequent windows of 50 trials. The horizontal dashed lines denote the estimated asymptotes T_d based on 500 trials (thick) and all trials (thin).

Mean Absolute Percentage Error (MAPE) A characteristic of the MPE is that positive and negative error values in individual trial times neutralise each other, such that this measure is a poor indicator of overall fit. Therefore we also report (table 3) the Mean Absolute Percentage Error (MAPE, StatSoft, 2001), which is computed as the average absolute error value, expressed as a percentage. However, when the distribution of errors is asymmetrical, the MAPE will not be minimal, even if the learning curve describes exactly the expected values.

The average MAPE for the full extrapolation is 35.7 percent (standard deviation 12.3 percent). The MAPE reflects that the average prediction of a single trial time is "off" by 35.7 percent. This indicates that single trial times on complex tasks are largely unpredictable.

The MAPE generally increases with task complexity. For the most complex task (task 3) the MAPE for the full extrapolation is 44.3 percent (table 3c). Table 3a reveals that this figure differs little from the MAPE in the described trials, which, on average, is 42.8 percent for task 3.

<u>Root of Mean Squared Errors (RMSE) and R^2</u>. We additionally report (appendix B) the commonly used RMSE and the percentage of variance accounted for (R^2). The table in appendix B reveals that both RMSE and R^2 decrease when applied to the extrapolations. Curiously, the decrease in RMSE would point to a better fit and the (dramatic) drop in R^2 would point to a worse fit. Considering this apparent paradox, one should recall that the interpretation of these measures depends on the assumption of stable additive Gaussian error, or, at least, on a stable and balanced distribution of errors, which is not justified when fitting a learning curve function to trial times. Both measures are based on squared errors, and therefore emphasise extreme error values. The latter occur frequently in trial time data whose probability distribution is skewed, not symmetric. Moreover, the mean and variance of trial times tend to decrease as a function of trial number Thus, both RMSE (the root of the mean of squared errors) and R^2 (the coefficient of determination) are inappropriate goodness-of-fit measures for the current application.

Table 3

Learning curve model: lack-of-fit measures[12].

A – Descriptive model (for all trials)					
Task	N[B] [13]	**MPE [%]**		**MAPE [%]**	
		Mean	**SD**	**Mean**	**SD**
1	8	-7.31	7.03	26.6	6.58
2	2	0.49	-	43.1	-
3	6	2.20	8.31	42.8	11.6
Total	16	**-2.77**	**8.32**	**34.7**	**11.9**

B – Extrapolated model (for trials 501-1000)					
Task	N	**MPE** [%]		**MAPE** [%]	
		Mean	**SD**	**Mean**	**SD**
1	8	-4.28	11.8	27.9	6.69
2	2	-2.45	-	37.5	-
3	6	5.30	7.50	40.4	8.47
Total	16	**-0.46**	**10.3**	**33.8**	**9.14**

C –Extrapolated model (for trials 501-All)					
Task	N	**MPE [%]**		**MAPE [%]**	
		Mean	**SD**	**Mean**	**SD**
1	8	-1.28	14.2	28.7	9.03
2	2	-1.11	-	38.1	-
3	6	12.9	10.9	44.3	13.1
Total	16	**4.07**	**13.7**	**35.7**	**12.3**

[12] Mean Percentage Error (MPE) and Mean Absolute Percentage Error (MAPE) in the model's prediction of trial times.

[13] Number of trainees over which Mean and Standard Deviation (SD) in the lack-of-fit measure is calculated.

Table 4

Learning curve model: Predicted training time, observed training time and relative error in the predicted training time for 16 trainees. (A) predictions over 500 trials (trials 501-1000); (B) predictions over on average 1800 trials (trials 501-total number of trials).

	A – Trials 501-1000			**B - Trials 501-total no. of trials**			
Trainee	**Pred. training time [min]**	**Obs. training time [min]**	**Rel. Error [%]**	**Pred. training time [min]**	**Obs. training time [min]**	**Rel. Error [%]**	**Total no. of trials**
1-1	81.8	72.3	13.2	471.0	418.3	12.6	3857
1-2	66.9	86.9	-23.0	328.0	429.9	-23.7	3327
1-3	80.1	92.3	-13.3	328.5	392.0	-16.2	2880
1-4	95.4	94.2	1.26	386.1	385.5	0.15	2921
1-5	97.0	97.7	-0.68	361.7	369.4	-2.08	2618
1-6	123.3	141.3	-12.7	351.2	370.2	-5.11	2112
1-7	171.1	159.9	7.0	398.5	332.6	19.8	1758
1-8	171.9	183.0	-6.1	304.3	291.8	4.26	1447
Mean (SD)	**111.0**	**116.0**	**-4.3 (11.8)**	**366.2**	**373.7**	**-1.28 (14.2)**	**2615**
2-1	136.8	136.9	-0.1	359.3	347.1	3.5	1971
2-2	159.5	167.6	-4.8	243.0	257.7	-5.7	1312
Mean	**148.1**	**152.2**	**-2.45**	**301.1**	**302.4**	**-1.1**	**1641**
3-1	151.2	142.7	6.0	696.3	625.4	11.3	3346
3-2	184.5	177.2	4.1	526.1	439.6	19.7	2282
3-3	188.5	186.2	1.2	546.0	480.5	13.6	2229
3-4	181.9	180.2	1.0	488.6	491.5	-0.6	2090
3-5	221.8	222.7	-0.4	361.7	348.2	3.9	1379
3-6	254.0	211.9	19.8	371.5	286.4	29.7	1273
Mean (SD)	**196.8**	**186.8**	**5.3 (7.5)**	**498.3**	**445.3**	**12.9 (10.9)**	**2100**
Mean (SD)	**147.3**	**147.9**	**-0.5 (10.3)**	**407.6**	**391.6**	**4.1 (13.7)**	**2300**

Training time predictions

In table 4 we tabulate the predictions of the model with respect to training time duration for individual trainees. In order to make these predictions, dead time estimate $\hat{T}_d$ and average trial time $<T>$ (equation 3) were calculated on the basis on the first 500 trials (which took on average 284 minutes of training time). On this basis we can use equation 3 to calculate the training time required for the next 500 trials (table 4a) and the training time required for all remaining trials (table 4b). From the predicted training time and the actual observed training time, the relative (percentage) error can be calculated. It should be noted that this relative error is equal to the previously discussed MPE, except for round off error (training times have been rounded off to the nearest 0.1 minute). For all tasks and all trainees together, the model predicted an average training time of 147.3 minutes (roughly 2.5 hours) for completion of the next 500 trials. The actual required training time was on average 147.9 minutes, a negligible difference. For all remaining trials beyond the first 500 trials, the model predicted that on average 407.6 minutes (6 hours, 48 minutes) would be required, whereas in fact an average of 391.8 minutes (6 hours, 32 minutes) was required.
The errors in predictions at the level of the individual trainee may be considerably larger, as reflected by the standard deviations in the relative error (these errors can be assumed to have a normal distribution). The most erroneous prediction is the prediction of the total training time of Trainee 3-6, the prediction being 371.5 minutes (roughly 6 hours, 11 minutes) for the completion of 2827 trials (total number of trials of Trainee 3-6 minus 500), whereas in fact only 286.4 minutes (4 hours, 46 minutes) were required, an absolute error of one hour and 25 minutes.

5 Discussion of results

Performance of the model

The results demonstrate the ability of a simple one-parameter function (a progressive average function), based on the linear rate model, to predict the fundamental trend in learning complex tasks, irrespective of the precise organisational structure of the task or its response requirements. The speed-based setting employed in this research allowed us to assume that skill level can be measured principally by the speed at which trials on the tasks are completed. The learning curve can thus be considered as the expected course (the trend) in the trial times for completion of the task.
The linear rate model predicts the expected value of trial times and derived quantities, such as the practice time needed to perform a certain number of trials. The model predicts a decreasing trend in trial times in which the expected value of each subsequent trial time is half the average of all preceding trial times plus half the dead time. We tested the latter prediction against data from 16 trainees and three different tasks of increasing complexity. Each of these tasks is known to require tens of hours of practice training before being mastered.
Evaluation of the learning curve model yields the following results:

- The Mean Percentage Error in the model's description of the complete learning curve (2300 trials on average, corresponding to on average 11.3 hours of training) of individual trainees is on average –2.77 percent.
- When, on the basis of the first 500 trial times, the second 500 trial times are predicted (corresponding to on average 2.5 hours of practice time), the Mean Percentage Error over those 500 trials is on average –0.46 percent.
- When, on the basis of the first 500 trial times, all remaining trial times are predicted (on average 1800 trials, corresponding to on average 7.2 hours of practice time), the Mean Percentage Error over those trials is on average 4.1 percent.

The trend in trial times is slightly better predictable when the task is less complex and the extrapolation period is shorter, as is reflected by the Mean Percentage Errors (MPE, table 3) for the distinct tasks and extrapolation periods. Moreover, the natural variation in trial times, as reflected by the Mean Absolute Percentage Error (MAPE, table 3), will increase as the task becomes more complex. In other words, the random trial-to-trial variations are larger with more complex tasks. The average MAPE for the full extrapolation is 35.7 percent (standard deviation 12.3 percent), which means that the average prediction of a single trial time is "off" by 35.7 percent. This indicates that single trial times on complex tasks are largely unpredictable by the model. As the MAPE predominantly represents random trial-to-trial variations, we argue that it is unlikely that this error will substantially decrease with more refined models (that is, with models based on more elaborate theories or with more parameters).

The irregularities in the individual learning curve that account for the Mean Percentage Error in individual extrapolations may be largely attributable to two sources: (1) The model is unable to predict extremely large trial times, which seem to occur more frequently during certain periods. Extremely large trial times are possibly due to temporary loss of skill in the time spans between training sessions. These extremely large trial times cause the model to underestimate the trial times, such that a negative Mean Percentage Error results. (2) The estimate of dead time T_d may be inaccurate. When the estimate of T_d is based on a lower number of trial times, the estimate will be more inaccurate. Generally, the estimates of T_d based on 500 trial times are slightly higher than the more accurate estimates based on 1000 or more trials, which causes overestimation of trial times and hence a positive Mean Percentage Error.

Estimating dead time

No tasks can be performed in an infinitely small amount of time. We denoted the incompressible part of a trial time by dead time T_d. The parameter has commonly been called an 'asymptote' (Newell & Rosenbloom, 1981, Logan, 1992, Speelman & Kirsner, 2001) and it has been interpreted as the theoretical minimum response time for a particular task determined by human processing limitations and/or mechanical limitations. In terms of the probability distribution of trial times, this parameter can be viewed as a position parameter, which defines the location of the lower bound of the probability density function of trial times (see e.g. Cousineau, Goodman & Shiffrin, 2002).

To determine this parameter accurately from a series or trial times $\{T_1, \ldots, T_n\}$ is not a trivial task. For example, Seibel (1963) investigated trial times for a complex pattern discrimination task (a '1023-alternatives task'). Seibel fitted a power function with dead time parameter T_d to nine learning curves, each consisting of thousands of trial times, using least-squares regression. In six out of those nine cases he found negative, implausible parameter values for T_d. On the basis of these results Seibel remarked (p 219-220): "Thus while the fitted function accounts for more than 93 percent of the variance of the data, the fits are statistically and theoretically not adequate".

In a comparative review of learning curve functions, Newell & Rosenbloom (1981) fitted the same power function with the dead time parameter T_d to trial times for 15 complex tasks of varying nature (including Seibel's data). Although their curve fitting procedure was a complicated mixture of computation and visual judgment, Newell & Rosenbloom (table 2.1, p. 24) found highly implausible T_d values of zero seconds for even the most complex tasks found in industry.

In a more recent comparative review by Heathcote, Brown & Mewhort (2000), the same power function was fitted to 17 published sets of empirical learning data (and numerous unpublished sets). Heathcote et al. used various combinations of least-squares regression. It was found that 87.7 percent of all estimates of asymptotic performance were physically implausible. This strengthened their arguments that a power function should be totally dismissed as the 'ubiquitous' description of learning curves.

Apparently, common techniques, such as those based on least-squares regression of the learning curve function fail to come up with reasonable estimates for the dead time parameter T_d. We argue that a primary reason for finding bad estimates for T_d with least-squares regression is an unsatisfied minimum condition for application of these techniques: a stable and recognized distribution of residual error.

In this research we presented an iterative method for estimating T_d, independent of the estimation of any other parameter and not reliant on the full distribution of residual errors. The method only assumes a distribution for the smallest residual error observed in a series of trials. By using a first estimate for T_d in the assumed learning curve function, we can approximate the random trial-to-trial errors. We subsequently obtain an improved estimate of T_d by fitting a Weibull distribution with specified scale and shape to the smallest error. This distribution is assumed for the smallest error, and is a consequence of the assumed learning mechanism, based on paired selection (Roessingh, Koenderink & Kappers, 2001b). More generally, the Weibull distribution is predicted by the instance theory (Logan, 1992). To further improve the estimate for T_d, the method can be iterated until it converges to a stable estimate.

We argue that this method for estimating dead time T_d is more robust than least-squares regression of the learning curve when the probability distribution of residual errors is only known partially. For the 16 learning curves studied in this experiment, application of the method resulted in plausible estimates.

6 Conclusions

The progressive average function has proved robust for extrapolation over long series of trial times (on average 1800 trials ~ 7.2 practice hours) of individual trainees who practiced with complex tasks.

The Mean Percentage Errors in individual predictions for the time needed to complete such a large number of trials are on average 4.1 percent with a standard deviation of 13.7 percent. The predictions do not depend on the precise organisational structure of the task or its response requirements. The progressive average function simply states that the expected value of each subsequent trial time is half the average of all preceding trial times plus half the dead time. The dead time parameter (‘asymptotic trial time’) causes the learning curve to saturate when practice time proceeds to infinity. For the purpose of estimating the dead time parameter from a series of trial times we presented a novel algorithm. Unlike usual techniques, such as least-squares regression of the learning curve, this algorithm provides plausible estimates for individual dead time.

The observed similar trend in all individual learning curves can be straightforwardly explained in terms of a simple cognitive strategy for learning a complex task from practice: In each task trial, the trainee compares the memory representation that controls his/her task execution with actual task execution. This leads to a new and likely superior memory representation of task execution. The latter is added to the set of existing memory representations. In a subsequent trial, a memory representation is randomly selected from this set.

This accumulative scheme predicts the observed trend in a series of trial times during a learning process and predicts the observed statistical properties of trial times. The explanation may be viewed as a simplified version of the instance theory of automaticity (Logan, 1988a, 1992), applicable to the acquisition of complex skills.

The authors thank Dr. John van Rooij (Royal Netherlands Army) for enabling the data collection, Dr. Stefan Louw and Dr. Harold Nefs for valuable remarks concerning the analysis. Finally, the authors are grateful to the sixteen participants who devoted their time to the experiments.

7 References

Cousineau, D., Goodman, V.W. & Shiffrin, R.M. (2002) Extending statistics of extremes to distributions varying in positions and scale and the implications for race models. To appear in Journal of Mathematical Psychology.

Donchin, E. (1989). The Learning Strategies Project, introductory remarks. Acta Psychologica, Vol. 71, 1-15.

Foss, M.A., Fabiani, A., Mané, A.M. & Donchin, E. (1989). Unsupervised practice; the performance of the control group. Acta Psychologica, Vol. 71, 23-51.

Frederiksen, J.R. & White, B.Y. (1989). An approach to training based upon principled task decomposition. Acta Psychologica, Vol. 71, 89-146.

Gopher, D., Weil, M. & Bareket, T. (1994). Transfer of skill from a computer game trainer to flight. Human Factors, 36(3), 387-405.

Hart, S.G. & Battiste, V. (1992). Field test of video game trainer. In: Proceedings of the Human Factors Society, 36th Annual Meeting.

Heathcote, A., Brown, S. & Mewhort, D.J.K. (2000). The Power Law repealed: The case for an exponential law of practice, Psychonomic Bulletin & Review, Vol. 7, 2, 185-207.

Logan, G.D. (1988). Towards an instance theory of automatization. Psychological Review, Vol. 95, No. 4, 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, Vol. 18, No.5, 883-914.

Mané, A. & Donchin, E. (1989). The Space Fortress Game. Acta Psychologica, Vol. 71, 17-22.

Newell, A. & Rosenbloom P. (1981). Mechanisms of skill acquisition and the Law of Practice. In J. R. Anderson (Ed), Cognitive Skills and their acquisition (pp. 1-55) Hillsdale N.J: Erlbaum Associates.

Roessingh J.J.M., Koenderink, J.J. & Kappers, A.M.L. (2001a). Complex skill acquisition and the linear rate model. Article submitted for publication, 2001.

Roessingh J.J.M., Koenderink, J.J. & Kappers, A.M.L. (2001b). Acquisition of complex manual control skills. Article submitted for publication, 2001.

Seibel. R. (1963). Discrimination reaction time for a 1,023-alternative task. Journal of Experimental Psychology, Vol. 66. No. 3, 215-226.

Singpurwalla, N.D. & Wong, M.T. (1983). Estimation of the failure rate-A survey of nonparametric methods. Part I: Non-Bayesian methods. Communications in Statistics - Theory and Methods, A12. 559-588.

Speelman, C.P. & Kirsner, K. (2001). Predicting transfer from training performance. Acta Psychologica, Vol. 108. 247-281.

Vidulich, M.A., McCoy, A.L. & Crabtree, M.S. (1995). Attentional control and situational awareness in a complex air combat simulation. Paper presented at AGARD Symposium on "Situation Awareness". Published in AGARD report CP-575. Neuilly-sur-Seine, France.

Appendix A: Summary of trial time data.

	ALL TRIALS				FIRST 500 TRIALS				SECOND 500 TRIALS (501-1000)			
Task-Subj.	**Max. trial time** [s]	**Min. trial time** [s]	**Mean trial time** [s]	**SD trial time** [s]	**Max. trial time** [s]	**Min. trial time** [s]	**Mean trial time** [s]	**SD trial times** [s]	**Max. trial time** [s]	**Min. trial time** [s]	**Mean trial-time** [s]	**SD trial times** [s]
1-1	112	5.2	8.5	5.2	112	6.4	15.3	10.3	39	5.7	8.7	3.0
1-2	98	4.4	9.6	6.3	98	5.4	12.7	9.0	73	5.3	10.4	7.4
1-3	120	4.7	11.0	7.2	120	6.1	16.1	13.5	57	5.6	11.1	5.1
1-4	502	5.9	11.4	11.9	502	8.3	20.5	26.4	40	7.0	11.3	3.3
1-5	393	5.6	12.1	10.3	393	6.9	19.1	20.3	53	6.4	11.7	5.3
1-6	129	6.6	16.8	12.0	129	9.2	26.4	16.6	74	7.7	17.0	9.2
1-7	198	7.4	21.0	15.2	198	11.9	34.0	22.3	58	9.3	19.2	6.3
1-8	204	6.5	25.4	19.6	204	9.1	38.5	25.8	101	7.1	22.0	10.9
M	**220**	**5.8**	**14.5**	**11.0**	**220**	**7.9**	**22.8**	**18.0**	**62**	**6.7**	**13.9**	**6.3**
SD	**149**	**1.0**	**6.0**	**4.8**	**149**	**2.1**	**9.3**	**6.7**	**20**	**1.3**	**4.8**	**2.8**
2-1	267	6.2	18.0	14.3	267	8.1	29.4	22.6	58	7.0	16.4	7.2
2-2	1198	5.9	27.2	42.2	1198	6.4	40.5	65.2	70	5.9	20.1	9.9
M	**733**	**6.1**	**22.6**	**28.2**	**733**	**7.3**	**34.9**	**43.9**	**64**	**6.5**	**18.3**	**8.5**
3-1	638	5.8	16.2	16.4	638	8.8	33.1	35.6	87	7.6	17.1	8.2
3-2	1399	4.6	22.4	47.9	1399	5.0	49.3	96.1	127	6.3	21.3	13.3
3-3	853	6.1	23.2	31.7	853	8.5	45.7	59.9	108	7.2	22.3	10.9
3-4	1098	5.8	24.9	30.9	1098	8.8	45.0	56.6	77	6.7	21.6	10.2
3-5	867	7.7	35.5	42.0	867	13.4	56.0	63.5	77	7.8	26.7	10.8
3-6	1433	7.3	38.5	66.3	1433	11.0	63.7	99.9	76	8.6	25.4	11.2
M	**1048**	**6.2**	**26.8**	**39.2**	**1048**	**9.2**	**48.8**	**68.6**	**92**	**7.4**	**22.4**	**10.7**
SD	**320**	**1.1**	**8.5**	**17.1**	**320**	**2.8**	**10.4**	**24.8**	**21.2**	**0.8**	**3.4**	**1.7**

Appendix B: Learning model: individual lack-of-fit measures.

	Lack-of-fit in described trials (All)				Lack-of-fit in extrapolated trials (501-1000)				Lack-of-fit in extrapolated trials (501-All)			
Task-Subj	**MPE** [%]	**MAPE** [%]	**RMSE** [s]	**100·R^2** [%]	**MPE** [%]	**MAPE** [%]	**RMSE** [s]	**100·R^2** [%]	**MPE** [%]	**MAPE** [%]	**RMSE** [s]	**100·R^2** [%]
1-1	0.34	19.4	3.52	55.5	13.2	28.1	3.13	3.04	12.6	26.5	2.72	7.37
1-2	-17.3	23.9	5.83	19.3	-23.0	24.4	7.64	20.4	-23.7	22.9	5.77	5.41
1-3	-12.6	23.7	6.17	32.3	-13.2	21.9	5.49	0.19	-16.2	20.8	4.83	2.29
1-4	4.12	21.8	6.35	74.0	1.27	18.6	3.62	2.38	0.15	17.5	3.18	14.7
1-5	-5.83	25.8	6.27	65.7	-0.68	29.9	5.34	0.02	-2.08	27.1	4.60	1.18
1-6	-10.6	33.4	10.3	32.2	-12.7	32.4	9.17	4.64	-5.11	32.9	7.89	8.07
1-7	-5.65	25.1	11.0	56.3	7.00	27.4	6.17	8.53	19.8	37.2	5.76	30.4
1-8	-10.9	39.4	15.8	47.7	-6.11	40.3	10.9	1.85	4.26	44.6	9.28	15.6
Mean	**-7.31**	**26.6**	**8.16**	**47.9**	**-4.28**	**27.9**	**6.43**	**5.12**	**-1.28**	**28.7**	**5.51**	**10.6**
SD	**7.03**	**6.58**	**3.94**	**18.7**	**11.8**	**6.69**	**2.66**	**6.74**	**14.2**	**9.03**	**2.22**	**9.53**
2-1	-2.05	37.2	10.4	48.6	-0.08	35.0	7.40	0.06	3.50	37.8	6.39	5.48
2-2	3.04	48.9	25.3	66.6	-4.82	40.0	9.82	1.67	-5.72	38.5	9.22	2.97
Mean	**0.49**	**43.1**	**17.8**	**57.6**	**-2.45**	**37.5**	**8.61**	**0.86**	**-1.11**	**38.1**	**7.81**	**4.22**
3-1	-2.80	28.4	10.49	60.0	6.00	33.5	7.97	9.60	11.3	33.6	5.71	14.0
3-2	17.4	61.1	28.2	66.3	4.12	51.2	12.9	9.27	19.7	58.6	9.18	23.1
3-3	5.29	40.2	18.7	65.6	1.23	37.4	11.2	1.53	13.6	41.8	8.13	17.7
3-4	-5.07	37.7	26.2	28.9	0.97	39.5	10.4	0.04	-0.6	36.0	8.61	6.79
3-5	1.12	37.7	23.7	68.1	-0.39	30.7	10.7	5.77	3.87	33.0	9.91	15.4
3-6	-2.76	51.6	61.0	16.9	19.8	50.0	11.2	13.4	29.7	62.6	11.1	24.8
Mean	**2.20**	**42.8**	**28.0**	**51.0**	**5.30**	**40.4**	**10.7**	**6.60**	**12.9**	**44.3**	**8.77**	**17.0**
SD	**8.31**	**11.6**	**17.3**	**22.3**	**7.50**	**8.47**	**1.60**	**5.14**	**10.9**	**13.1**	**1.82**	**6.54**