COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 42 (2018, Suppl. 3) 783–808

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12599

Predictive Movements and Human Reinforcement Learning of Sequential Action

Roy de Kleijn,[a] George Kachergis,[b] Bernhard Hommel[a]

[a]*Cognitive Psychology Unit, Leiden University*
[b]*Department of Artificial Intelligence, Radboud University*

Received 24 February 2017; received in revised form 19 December 2017; accepted 22 January 2018

Abstract

Sequential action makes up the bulk of human daily activity, and yet much remains unknown about how people learn such actions. In one motor learning paradigm, the serial reaction time (SRT) task, people are taught a consistent sequence of button presses by cueing them with the next target response. However, the SRT task only records keypress response times to a cued target, and thus it cannot reveal the full time-course of motion, including predictive movements. This paper describes a mouse movement trajectory SRT task in which the cursor must be moved to a cued location. We replicated keypress SRT results, but also found that predictive movement—before the next cue appears—increased during the experiment. Moreover, trajectory analyses revealed that people developed a centering strategy under uncertainty. In a second experiment, we made prediction explicit, no longer cueing targets. Thus, participants had to explore the response alternatives and learn via reinforcement, receiving rewards and penalties for correct and incorrect actions, respectively. Participants were not told whether the sequence of stimuli was deterministic, nor if it would repeat, nor how long it was. Given the difficulty of the task, it is unsurprising that some learners performed poorly. However, many learners performed remarkably well, and some acquired the full 10-item sequence within 10 repetitions. Comparing the high- and low-performers' detailed results in this reinforcement learning (RL) task with the first experiment's cued trajectory SRT task, we found similarities between the two tasks, suggesting that the effects in Experiment 1 are due to predictive, rather than reactive processes. Finally, we found that two standard model-free reinforcement learning models fit the high-performing participants, while the four low-performing participants provide better fit with a simple negative recency bias model.

Correspondence should be sent to Roy de Kleijn, Cognitive Psychology Unit, Leiden University, Wassenaarseweg 52, 2333AK Leiden, The Netherlands. E-mail: kleijnrde@fsw.leidenuniv.nl

Keywords: Sequence learning; Serial reaction time task; Sequential action; Reinforcement learning; Movement trajectory; Implicit motor learning

1. Introduction

Most daily human behaviors are learned sequential actions: from walking, cooking, and cleaning to speaking and writing. Consequently, sequence learning has been studied in different contexts ranging from implicit sequence learning (Boyer, Destrebecqz, & Cleeremans, 2005; Cleeremans & McClelland, 1991; Nissen & Bullemer, 1987; Stadler, 1992) to language acquisition (Elman, 1990; Saffran, Newport, & Aslin, 1996), typing (Fendrick, 1937; Gentner, LaRochelle, & Grudin, 1988), and manual everyday actions (Botvinick & Plaut, 2004; Cooper & Shallice, 2000). In implicit learning research, an important paradigm has been the serial reaction time (SRT) task, which requires participants to press one of four buttons when cued by a corresponding light, in a sequence that repeats—unbeknownst to learners—every 10 presses (Nissen & Bullemer, 1987). Subjects trained on this repeating sequence developed faster reaction times (RTs) over the course of training, as compared to a control group responding to a random sequence of stimuli. The SRT paradigm has been cited as evidence for implicit learning, as subjects experiencing the repeating sequence, despite showing faster RTs over time, report no explicit knowledge of the sequence when debriefed afterwards. However, performance does suffer somewhat when participants must simultaneously perform a second task (Nissen & Bullemer, 1987), suggesting that learning in the SRT task does require some attentional resources or effort. The role of attention in the SRT task was further studied in Fu, Fu, and Dienes (2008), who demonstrated that reward motivation can improve the development of awareness of the sequence. Fu et al. (2008) reasoned that reward motivation regulates the amount of attention paid toward the stimuli, which in turn facilitates sequence learning. Additionally, Willingham, Nissen, and Bullemer (1989) found that some participants achieved a degree of declarative knowledge after a fixed training period in the SRT task, and that additional training resulted in more explicit knowledge for many subjects, if not all. On balance, it seems that the SRT task is neither wholly implicit nor wholly explicit, but in the literature the speed-up observed over time is often regarded as evidence for implicit learning (Seger, 1994), while declarative sequence knowledge or performance on a generation task is often regarded as explicit knowledge.

The dissociation of implicit and explicit processes facilitating sequence learning remains a topic of debate, yet learning remains robust under high degrees of noise and complex structure in the sequences (Cleeremans & McClelland, 1991). Complex action sequences are not simple stimulus-response chains, but rather require representing sequential context in order to learn (Lashley, 1951). Moreover, human behavior is often thought of as predictive—indeed, many models of sequential learning operate on a prediction-based error signal (Botvinick & Plaut, 2004; Kachergis, Wyatt, O'Reilly, de Kleijn & Hommel, 2014). Thus, it is problematic that the discrete button-presses in the SRT paradigm cannot distinguish an *anticipatory* response due to correctly predicting the stimulus (or a slow response due to an incorrect prediction) from *reactive* responses (although perhaps pre-potentiated) based on the cue (Marcus, Karatekin, & Markiewicz, 2006). Truly predictive responses—that is, those made before the next response is cued (500 ms after the previous response)—are not valid, allowed responses in the SRT paradigm. These shortcomings of discrete button-press responses have been discussed and addressed previously. For example, Moisello et al. (2009) employed an arm-reaching paradigm to distinguish two components of motor responses: the time between stimulus appearance and response initiation thought to reflect declarative knowledge, and the response execution time thought to reflect implicit knowledge, which would not be possible in a discrete button-press paradigm. And, in another study, Marcus et al. (2006) combined discrete button-press responses with eye tracking to investigate anticipatory eye movements reflecting stimulus prediction.

In this paper, we describe two modifications of the SRT paradigm that allow us to naturally investigate both predictive and reactive responding in human sequence learning. In Experiment 1, recognizing that actions are continuous movements that can reveal the underlying dynamics of the cognitive processes driving them (Spivey & Dale, 2006), we used a mouse-tracking adaptation of the SRT task in which spatial locations are both stimuli and response options (Kachergis, Berends, de Kleijn, & Hommel, 2014a, b). By tracking their movement before and after the next target is cued, we investigated changes in predictive versus cued responding over the course of the experiment (Tubau, Hommel, & López-Moliner, 2007). Using this trajectory SRT paradigm, we replicated the overall Nissen and Bullemer (1987) RT results; moreover, we show sequential context effects—predictive bends in response trajectories—along with different movement dynamics pre- and post-cue.

In many implicit learning tasks such as artificial language learning and the SRT paradigm, learning is dependent on recognizing some statistically reliable sequential structure in stimuli that are not under the learner's control. However, everyday human action learning is often not characterized by processing a steady stream of stimuli, but by exploring the environment (i.e., choosing actions) and receiving positive and negative feedback. Prediction is thus an essential element of reinforcement learning (RL), which is a well-established paradigm in the field of machine learning (Sutton & Barto, 1998) that was originally motivated by much earlier behaviorist stimulus-response learning studies (Skinner, 1950).

Although reinforcement learning is now a robust subfield of machine learning with applications to AI and robotics, there is evidence that similar processes play a role in human learning. For instance, the error-related negativity (ERN) event-related potential (Falkenstein, Hohnsbein, Hoormann, & Blanke, 1991; Gehring, 1992) has been studied extensively as a component of error processing. The ERN originates in the brain whenever task-relevant errors are committed. Holroyd and Coles (2002) link the ERN to the mesencephalic dopamine system and propose it is the result of a negative reinforcement signal which it conveys to the anterior cingulate cortex. A recent brain imaging study found neural correlates for prediction error signals that correspond to those in some RL models (Bornstein & Daw, 2012). More recently, the reinforcement learning approach has demonstrated its value to artificial intelligence (Mnih et al., 2015; Silver et al., 2016), cognitive science (Dezfouli & Balleine, 2012; Gureckis & Love, 2009), and neuroscience (Averbeck & Costa, 2017; Frank & Badre, 2012) communities.

RL paradigms allow learning agents to interact with a task solely through observations, actions, and rewards. The rewards validate the actions, without the need for explicit cueing or other forms of instruction. Thus, learning is exploratory and accomplished via trial and error. In Experiment 2, we further modified the trajectory SRT paradigm by not cueing responses at all: Participants had to explore response alternatives until the correct one was found, receiving feedback (negative or positive points) at each response. In this aspect, Experiment 2 is a purely predictive version of Experiment 1, in which participants could employ either reactive or predictive strategies. We investigated sequence learning in this RL SRT paradigm and found correspondences between successful learners in this paradigm and in the reactive SRT paradigm in Experiment 1. Using the RL paradigm allowed us to study the effect of rewards on sequence acquisition in more detail, yielding not only response times but also errors over time. Thus, this study adapted the trajectory SRT task to allow for free movement and limited instruction, allowing learners to explore and learn from trial and error.

In addition, we attempted to capture human performance and error patterns using reinforcement learning models. Due to the relatively simple nature of the task, we investigated if simple (i.e., model-free) RL models were sufficient to learn the repeating sequence by trial and error. We assessed the RL data both in terms of earlier SRT data and in comparison to three standard RL models. Overall, this study provides insights into prediction error-driven learning of sequential action learning.

2. Experiment 1

The purpose of the first experiment was to use the trajectory SRT paradigm to replicate earlier findings by Nissen and Bullemer (1987). This study used four stimuli in a recurring sequence of length 10, horizontally displayed on a screen. Designating the stimulus positions from left to right as numbers, the original sequence read 4-2-3-1-3-2-4-3-2-1. To fit the trajectory paradigm the sequence was mapped to a square, left-to-right and top-to-bottom (i.e., 1 = top left, 2 = top right, 3 = bottom left, and 4 = bottom right). Participants moved the mouse from one stimulus position to the next, corresponding to the sequence. We tested two groups of participants, one trained on the recurring sequence and the other trained on a random sequence. After 10 blocks of training, participants completed a generating task. This task consisted of the same basic test conditions, except participants were asked to predict the sequence instead of following it.

Nissen and Bullemer (1987) originally found participants showing improved performance within the first block of training. The authors demonstrated that performance suffered under dual-task conditions and varied as a function of serial position in a pattern suggesting that learners were chunking the sequence into two pieces. In total, the study's results suggest that attention to the sequence is crucial for both implicit and explicit sequence learning, but that improved performance is not critically dependent on awareness of the sequence. For the purpose of Experiment 1 only, the initial experiment was replicated. We expected to replicate the basic improvement of performance, as well as the chunking pattern that was observed. Like Willingham et al. (1989), we included a final generation task, in which participants were asked to reproduce any action sequence they felt they had learned during training.

2.1. Methods

2.1.1. Participants

Participants in this experiment were 22 Leiden University undergraduate students who participated in exchange for 3.5 euros or 1 course credit.

2.1.2. Apparatus and materials

The experiment was performed on a computer with a 21" monitor with a 1,024 × 768 resolution and a 60 Hz refresh rate. Participants used a mouse to move the cursor. The experiment was programmed in Python with the PyGame library, and cursor position was sampled at every screen refresh. Four centimeters of mouse movement resulted in 1,000 pixels cursor displacement, and mouse cursor acceleration was disabled, as is suggested by Fischer and Hartmann (2014).

The stimulus display consisted of four red squares (location 1 = upper left, 2 = upper right, 3 = lower left, 4 = lower right), displayed continuously. Each stimulus was an 80 × 80 pixel square, separated by 440 pixels of white space. From center to center, each diagonal movement had a length of 735 pixels (2.94 cm mouse displacement), and each horizontal or vertical movement had a length of 520 pixels (2.08 cm mouse displacement).

Responses were given by moving the mouse cursor to the target, thereby touching it. This is in contrast to mouse-click responses, another frequently employed technique. For the current paradigm, in which longer sequences could be planned, and motion could be optimized for continuous movement, mouse-over responses were chosen.

2.1.3. Procedure

Participants were alternately assigned to one of the two between-subjects conditions according to the order they signed up. In the NB87 sequence condition, participants were given a repeating sequence of 10 locations corresponding to the Nissen and Bullemer (1987) sequence (4-2-3-1-3-2-4-3-2-1). In the random sequence condition, participants followed a randomly generated movement sequence without repetitions (i.e., staying at the same location).

The first part of the experiment consisted of a training phase. Participants were told to quickly and accurately move the mouse cursor to whichever square turned green. After arriving at the highlighted stimulus (i.e., the moment the mouse cursor first touched any part of the stimulus square), the response was recorded, and another stimulus was highlighted after a 500 ms ISI. In the NB87 condition, this would be the next stimulus in the sequence. In the random condition, a random stimulus turned green without repeating the current location. If the wrong stimulus was touched, no feedback was given, but the target remained green until it was touched. Participants were allowed to take as long as they liked (i.e., there was no timeout), but they were instructed to move as fast and accurately as possible. Participants completed 80 training trials, each of which contained a series of 10 locations, leading to a total of 800 movements. Participants were given a rest break every 20 training trials. Following the training phase, participants were asked to try to reproduce any sequence they had learned.

After the training phase, participants were given a generating task similar to the training phase. In the generating task, participants were asked to predict where they thought the stimulus would appear and move the mouse to that square. In other words, they were asked to complete the sequence without being cued. A correct prediction would cause no color change while an error would cause the correct continuation of the sequence to appear in green, and participants were to move to the next location.

2.2. Results

2.2.1. Response times

Data were analyzed from the 22 participants (11 per condition) that completed the experiment. Median movement time to a target was 1,040 ms (*SD*: 1,776). Of 17,578 target arrival times, 84 were removed for being slower than 2,816 ms (median + *SD*). Each subject's median RT for correct movements on each block was computed. Fig. 1a shows the mean of median RTs by block for the two conditions. Participants in both conditions got faster over the course of the experiment, but participants in the NB87 sequence condition improved more than those in the random condition, replicating the Nissen and Bullemer (1987) speedup. There was a 25% reduction in reaction time over the course of training. These data were analyzed by a two-way analysis of variance, which indicated significant main effects of condition ($F(1, 20) = 31.3$, $p < .001$) and block ($F(7, 168) = 6.3$, $p < .05$), and a significant interaction effect ($F(7, 210) = 14.7$, $p < .01$) between the two.

The accuracy data are shown in Fig. 1b. Accuracy was high across training blocks, although it dropped over time in the NB87 group, particularly after the first three blocks of training. A two-way analysis of variance confirmed a significant main effect of group ($F(1, 20) = 36.7$, $p < .001$) and a significant interaction effect ($F(9, 210) = 14.1$, $p < .001$). These results are evidence of sequence learning, replicating the Nissen and Bullemer (1987) keypress-based results. However, in the NB87 condition, both accuracy and RT dropped over time. In the NB87 condition, faster median hit RTs on a training block had a significant negative correlation with the number of errors in that block (for the 67 of 110 blocks containing errors; $r = -.56$, $t(65) = -5.48$, $p < .001$), showing a speed-accuracy tradeoff. This was not present in the Nissen and Bullemer (1987) results, but it can be explained through the difference in response execution. Key-presses are intermittent and can only be made in response

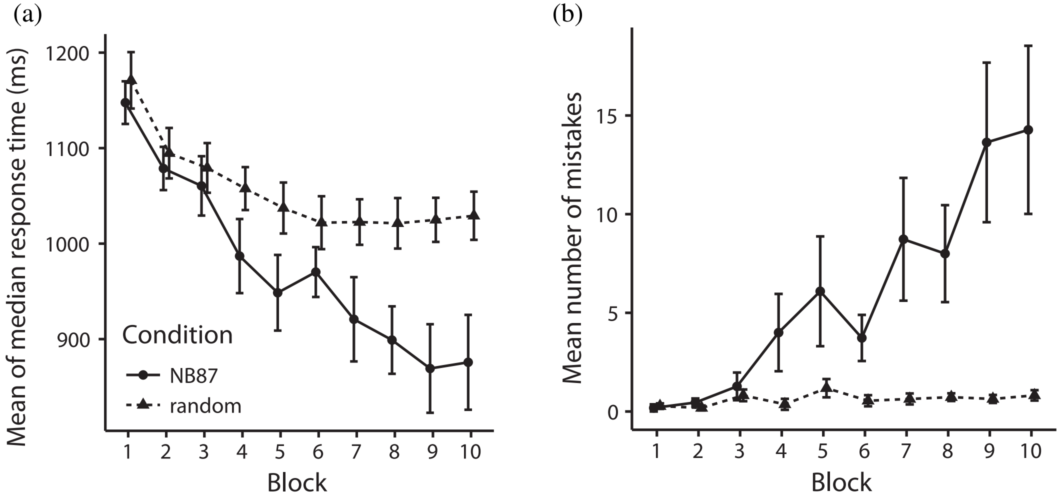

Fig. 1. Experiment 1 training phase RTs and error rates by block. (a) Mean of median RTs by block show that both conditions sped up over the course of Experiment 1, but that NB87 improved more. Error bars show +/−1SE. (b) Mean number of errors by block shows only the NB87 participants made an increasing number of errors. Error bars show +/−1SE.

to a stimulus (pre-stimulus responses were not recorded), while mouse movements are continuous and made constantly. This speed-accuracy trade-off is likely due to the trajectory SRT paradigm encouraging prediction, allowing participants to move freely while performing the experiment.

A two-way ANOVA with block as between- and serial position as within-subject factors, which showed significant main effects for block ($F(9, 210) = 32.3$, $p < .001$ and serial position ($F(9, 100) = 10.2$, $p < .01$). To determine whether participants became faster at the entire sequence or rather learned some chunks better than others, mean RT was plotted for each serial position, shown in Fig. 2. Similar to the Nissen and Bullemer (1987) results, RTs on the second, fifth, and eighth serial positions are slow, which may indicate that participants chunk the full sequence into two small, well-learned pieces.

Performance on the generating task was poor, as participants on average did not manage to reproduce the sequence without making many errors, as shown in Table 1. This indicates that, although training performance showed evidence of sequence learning, participants were not explicitly aware of the sequence. It is possible that participants would eventually be able to reproduce the sequence if training were extended, as in Willingham et al. (1989). Nissen and Bullemer (1987) originally found that participants were able to score around 80% correct on the generating task after two blocks of 10 trials. However, it should be noted that Nissen and Bullemer (1987) interleaved training and generation blocks, which could have produced a practice effect. Although the current study only required participants to complete one block of 10 trials during the generating task, participants did not show any improvement during the task.

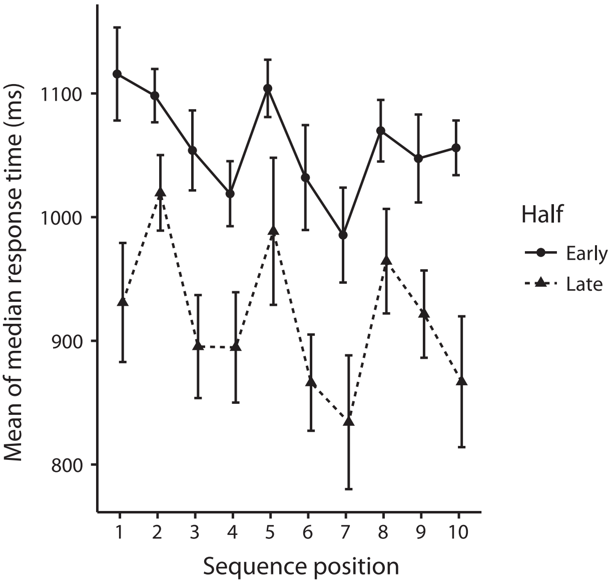

Fig. 2. Mean of median RT by sequence position during the early and late halves of training. Bars show +/−1SE.

Table 1
Generating task performance by NB87 participants (chance would be 6.67 errors on average)

Sequence Position	Hit RT	Average Errors
1	1560	5.64
2	1577	5.70
3	1716	5.80
4	1541	6.70
5	1574	7.09
6	1448	5.70
7	1482	5.50
8	1427	5.10
9	1426	5.33
10	1334	5.18

2.2.2. Trajectory results

The continuous nature of mouse movements allowed participants to actively move toward the next target before it appeared. Indeed, an analysis of distance to target at target onset, shown in Fig. 3, shows that participants in the NB87 condition increasingly move toward the next target in the 500 ms interval before the next target becomes highlighted. This shows that participants in the NB87 condition are correctly predicting the next target location and already moving toward it before the next cue appears, with an interaction effect between Condition and Block, $F(9, 180) = 4.21$, $p < .001$. Note that due to the structure of the sequence, the average distance between targets is slightly

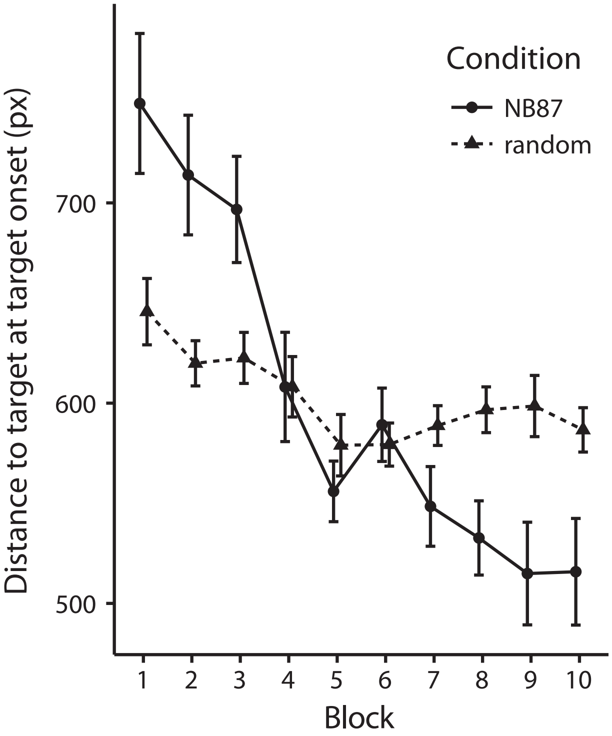

Fig. 3. Initial distance to target at target onset. Smaller values indicate movement toward the next target during the ISI (before the stimulus was visible). It is clear that participants in the NB87 condition show increased prediction over time. Error bars show +/−1 SE.

smaller in the random condition (591 pixels) than in the NB87 condition (605 pixels), which could partly explain the observed difference in early blocks.

Fig. 4 shows an example of mouse movements during a characteristic trial from each condition. Participants in the random condition (e.g., Fig. 4 left) tend to re-center the cursor after hitting a target, during the 500 ms ISI. This strategy is not unreasonable under conditions of uncertainty, as it minimizes the distance to potential targets, and the next target cannot be predicted in the random condition. Centering behavior is shown in Fig. 5a, and it is defined as the proportion of time spent in the center 100 × 100 pixels of the screen between reaching the previous target and current target reached. Instead of the binary distinction between reactive and predictive movement as used by Dale, Duran, and Morehead (2012), we chose to use their continuous measure of prediction to assess the magnitude of prediction. As the experiment progresses, participants in the random condition adopt a centering strategy that minimizes distance to potential targets, while participants in the predictable NB87 condition do not show this behavior. Participants in the random condition spend an increasingly larger proportion of time in the center of the screen compared to NB87 participants, $F(9, 180) = 2.51$, $p = .010$ for the interaction between block and condition. Similar centering behavior has been reported, but not quantified in the current context by Duran and Dale (2009), and Dale et al. (2012).

Interestingly, not all participants in the random condition display this centering strategy, as evidenced by the large standard errors, especially in the final half of the experiment. Instead, participants seem to employ either a non-centering strategy or a centering strategy in which they spend almost 25% of the ISI in the center of the screen.

With learning, targets are predictable in the NB87 sequence condition; thus, participants are expected to show faster reaction times (RTs) as training proceeds.

The NB87 sequence, 4-2-3-1-3-2-4-3-2-1, contains only one identical transition (3-2, a diagonal movement), although other movements are isomorphic (e.g., 4-2 and 3-1). We examined the development of sequential context effects—deflections in response trajectory caused by the prior or subsequent location—by plotting the average trajectories for the isomorphic movements: 4-2 versus 3-1. In the experiment, these movements are vertical, and we were interested in investigating the average deflections from the direct path from one stimulus center to another. We binned the movement time, starting at ISI onset, in bins of size 20, and averaged and plotted their deviation using local regression smoothing from the direct path (y-axis) over time (x-axis) in Fig. 6, split by condition, and for each half of training. Note that movement time reflects time from ISI onset, whereas RT reflects time from cue onset. Early in training, some centering behavior is apparent in both conditions, most notably in the 4-2 movement. This movement also clearly shows the absence of centering behavior late in training for the NB87 condition. The 4-2

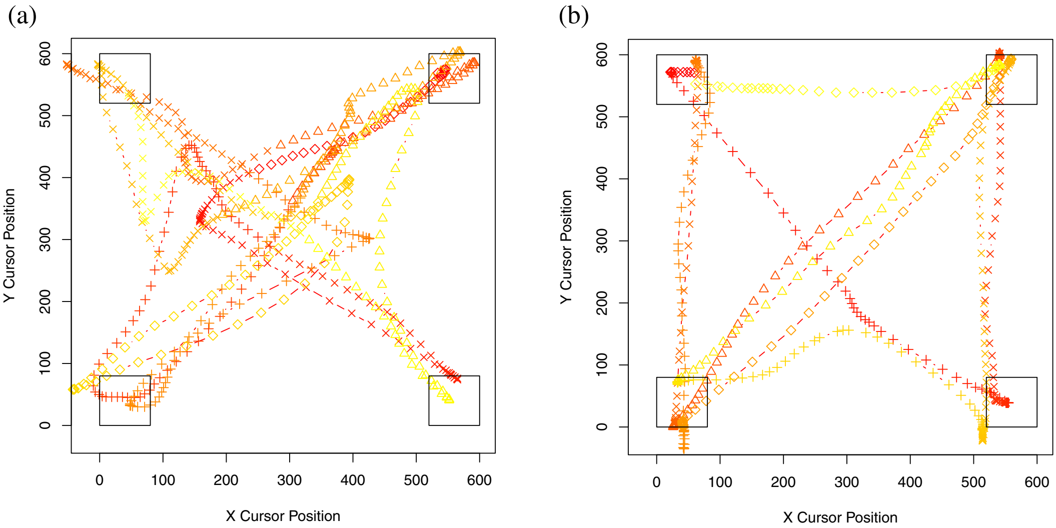

Fig. 4. Characteristic movements in one trial from each condition. (a) One trial from one participant in the random condition, in which the next location was chosen at random, without repeats. All 11 random participants adopted a similar strategy of re-centering the cursor after each response. This is optimal in the sense that it was impossible to know which location will be highlighted next (t_0 = red, t_{end} = yellow). (b) A characteristic trial of a participant's movements during the NB87 sequence, beginning at location 4 (lower right) and ending at location 1 (upper left). These isomorphic trajectories can be compared for context effects. Only four NB87 participants showed centering movements in the last half of training.

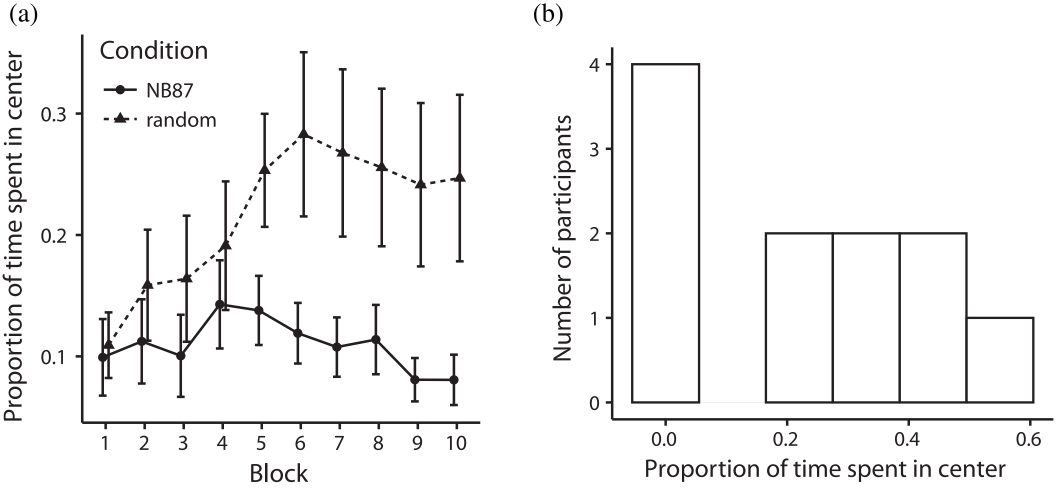

Fig. 5. Centering behavior during the ISI. (a) Proportion of time spent in the center of the screen, defined as a 100 × 100 pixel square in the center of the screen. Centering behavior in the random condition is clearly visible. Error bars show +/-1SE. (b) Distribution of centering behavior for the last half of the experiment for the random condition. Two groups of participants can be identified: those who center during the ISI and those who do not.

movement also shows participants tended to move toward the left after completing the movement. As the next target in the sequence is 3, which is situated to the bottom left of the current target, this indicates they were beginning to move toward the subsequent target. These trajectory analyses corroborate that NB87 participants were making increasingly predictive movements, bending toward the next stimulus position based on their contextual knowledge.

2.3. *Discussion*

In summary, Experiment 1 replicated the speed-up results from the Nissen and Bullemer (1987) serial button-pressing task with a mouse-trajectory version of the task, showing that participants learn regularities in the stimulus stream and exhibit speeded responding, even though they are bad at explicitly reproducing the sequence. We have also demonstrated the advantage of the trajectory-tracking SRT task: Because participants can move during the interstimulus interval—before the next cue has appeared—we can distinguish predictive movements (toward the correct next stimulus) from post-cue speedups. Indeed, we found that participants in the NB87 sequence condition made an increasingly large proportion of their movement during the 500 ms pre-cue interval. Looking at the RTs for each sequence position, we found that they correlate with the observations made by Nissen and Bullemer (1987), suggesting that the sequence is learned in chunks. Several chunk-based models exist that use such mechanisms to aid sequence parsing and acquisition, such as MDLChunker (Robinet, Lemaire, & Gordon, 2011), PARSER

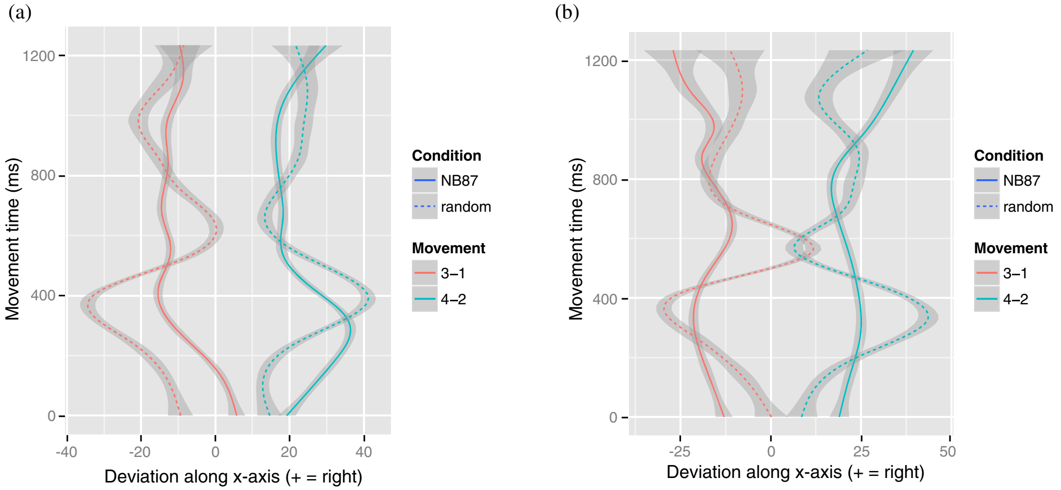

Fig. 6. Averaged trajectories for vertical movements 4-2 and 3-1 between 0 and 1,200 ms after ISI onset. (a) Horizontal deviation during movement (i.e., over time) in early training. Both conditions' trajectories show some centering behavior, bending toward the middle (i.e., right for 3-1, left for 4-2). NB87 trajectories show less deviation. (b) Horizontal deviation during movement in late training. The random condition shows more centering behavior, while the NB87 trajectories show little variation except at the end of the movements when they diverge, showing prediction of the subsequent stimulus.

(Perruchet & Vinter, 1998), and TRACX (French, Addyman, & Mareschal, 2011). Also, we found centering behavior similar to Dale et al. (2012) in some of the participants, with others either waiting for a cue or actively moving toward the next target. This bimodality of behavior has been reported in another predictive paradigm by Bruhn, Huette, and Spivey (2014), which shows some participants actively engaging in predictive behavior, where others engaged in a reactive "wait and see" manner. However, in addition to the study by Dale et al. (2012), we compared centering behavior between the random and NB87 condition, showing that participants in the random condition show significantly more centering behavior, which can be explained by uncertainty in prediction. Having established that prediction plays a role in the speed-up seen in the SRT-trajectory paradigm, in Experiment 2 we made prediction the essential goal of the task, requiring learners to move to the next location without a cue, and only giving feedback upon making a response.

3. Experiment 2

The results of Experiment 1 show that spatial sequences can be learned through cued learning, replicating a huge body of literature on the SRT task introduced by Nissen and Bullemer (1987). However, sequence learning in everyday action can hardly be considered cued. Instead, humans are in constant interaction with their environment, exploring it and receiving positive or negative feedback on their taken actions. In Experiment 2, we adapted the paradigm of the trajectory SRT into an exploration paradigm in which participants actively try out the alternative options and receive feedback (reinforcement or punishment). More specifically, the goal of Experiment 2 was to examine reinforcement learning within the trajectory SRT paradigm and to compare human performance to basic baseline models. The trajectory SRT task was adapted to no longer cue participants with the next target position, forcing them to instead explore the response alternatives until the correct one was found. Moving the mouse cursor from the previous target to another response alternative resulted in a reward (+1) or penalty (−1) that was accumulated throughout the experiment and displayed continuously. Upon reaching a valid target, it would change color to green, add to the score by +1, and allow the participant to continue exploring. Reaching for an invalid target caused it to change to red, subtract from the score by 1, while the cursor was relocated to the previously occupied target, effectively resetting the participant's progress. Target validity was determined by a recurrent sequence, taken from the Nissen and Bullemer (1987) study, and adapted to fit the trajectory SRT paradigm. Designating the stimuli as numbers from left to right, top to bottom, the sequence read 4-2-3-1-3-2-4-3-2-1.

3.1. Methods

3.1.1. Participants

Participants in this experiment were 13 Leiden University students and employees (age: M = 23.9, SD = 6.4) who participated in exchange for 3.5 euros or for course credit.

3.1.2. Procedure

Participants were instructed that they would be presented with four target squares in the corners of the screen which they were to explore by moving the mouse, each time resulting in either a gain or loss of one point. Participants were told to try to maximize their score, which was displayed continuously at the top of the screen. Unbeknownst to the participants, only one of the four targets would be valid at any given moment, but all were colored blue, so the target could not be visually distinguished. Upon reaching a valid target, its color would change to green momentarily and the score would increase by one. The participant would be able to continue exploring for the next target. Arriving at an invalid target caused it to change to red momentarily and the score was decreased by one, while the cursor was relocated to the previously occupied target. Thus, although there were no instructions explicitly indicating it, participants likely inferred that they had chosen the incorrect stimulus, and should choose one of the remaining two—if they also assumed the same target was never repeated immediately, which was true. In the absence of a previous target (i.e., at the beginning of the experiment or after a rest break), the cursor was moved back to the middle of the screen.

Unbeknownst to the participants, each trial consisted of a series of 10 targets (labeled 1-4 left-to-right and top-to-bottom: 4-2-3-1-3-2-4-3-2-1) that repeated continuously, with no indication where one trial stopped and the next began. Participants completed eight blocks of 10 such trials, with a short rest break after every two blocks (i.e., 200 correct movements). A participant who somehow knew the sequence before entering the experiment and never made an error would therefore make 800 movements to valid targets, receiving a theoretical maximum of 800 points. At worst, a participant with no memory of even the previous target she had tried may make an infinite number of errors and may never finish the experiment. Assuming enough memory to not repeat the same invalid target more than once when seeking each target (i.e., an elimination strategy), a participant using this elimination strategy would expect on average to score 0 points, as the expected value (EV) of completing one movement successfully is 0.[1] Note that participants were not told that there was a single deterministic sequence, let alone details such as how long the sequence was.

3.2. *Results*

The data from all 13 participants were analyzed. Fig. 7 shows a histogram of the final score achieved by each participant. Note that the distribution looks bimodal,[2] with four participants collecting less than 300 points and all but one of the rest accumulating more than 500 points each. Given the bimodal score distribution, a median split was used to divide the participants into high-performing (≥ 526; 7 people) and low-performing (< 526; 6 people) groups. In the high-scoring group, participants achieved almost flawless performance after only approximately 30 trials, with a final mean score of 652 (max: 725), while the low-scoring group only gradually increased their score (final mean score: 287). The remaining analyses were carried out for each group in an attempt to understand the great variability in performance—and the impressive success of the high-scoring group.

3.2.1. *Response times*

The overall median response time (RT) for all stimulus arrivals was 1,401 s (*SD*: 4,980). Of 10,400 correct target arrival times (median: 1,078 ms, *SD*: 2,216), 317 (3%) were trimmed for being too slow (median + 2 · *SD*). Of the 4,117 incorrect stimulus arrival times (median: 2,397 ms, *SD*: 8,401), 100 were trimmed for being too slow (2.4%). Each subject's median RT for correct and incorrect movements was computed for each 10-trial block. Fig. 8 shows the mean of subjects' median correct and incorrect RTs over the experiment, split into high- and low-performing group. RTs for correct movements improve in both groups during the first few blocks, but the high-scoring group speeds up more than the low-scoring group. Fig. 8 also shows that the rare incorrect RTs for the high-performing group get slower over the course of the experiment, whereas the low-performing group's incorrect RTs only increase a bit. The strikingly slow errors of high-performing participants, compared to errors that are barely slower than correct movements for the low performers may indicate a different mode

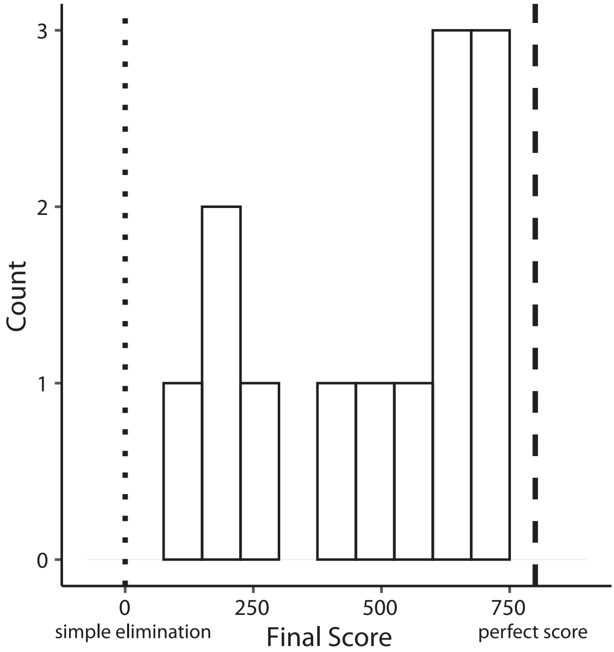

Fig. 7. The histogram of participants' final scores after completing 80 sequence repetitions (800 targets) shows a bimodal distribution (lines: elimination strategy EV = 0; perfect knowledge EV=800).

of behavior. A possible explanation is that low performers are simply not trying to learn a sequence, or they do not expect it to be deterministic, whereas high performers explicitly learn the sequence, and when they are uncertain they must pause to try to recall the next target.

3.2.2. Accuracy

The mean number of errors made over the entire experiment was 19.8 (*SD* = 21.3) for the high-scoring group and 63.5 (*SD* = 11.9) for the low-scoring group. Over time, the number of errors decreased, especially for the high-scoring group. Examining the errors made by each group of participants according to where they were in the sequence revealed that for both groups the fifth stimulus was particularly challenging. This is reflected in the mean number of errors for each group (see Fig. 9b, as well as in the mean RT to the target by sequence position; see Fig. 9a).

3.2.3. Comparison to Experiment 1

The pattern we observe in the accuracy and response time data bears some resemblance to the pattern observed in Experiment 1, despite the use of cues in that experiment. Although the RL SRT task in Experiment 2 was fundamentally different from the cued SRT task in Experiment 2, the same sequence was used in both experiments. We can therefore scale ([−1,1]) the by-sequence position mean response times from

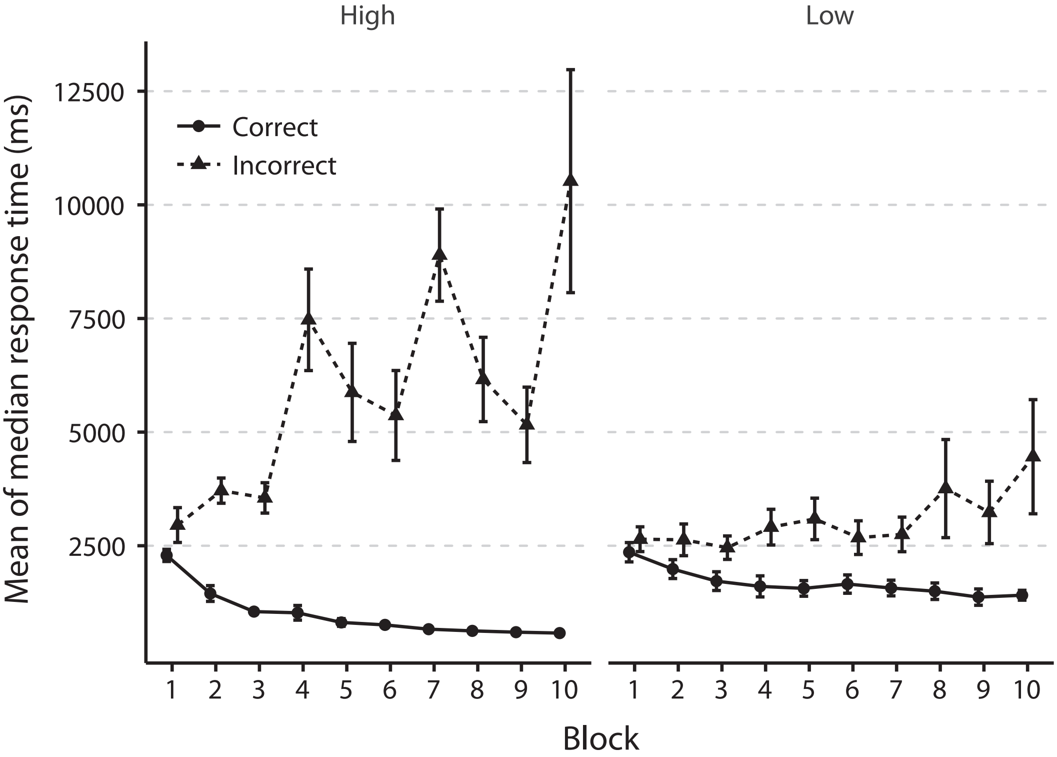

Fig. 8. The mean of subjects' median correct RTs by block shows that high-performers' (left panel) RTs improved more than the low-performers' (right panel) RTs over training. The mean of subjects' median incorrect RTs by block shows that the high-performing group's incorrect RTs actually increased, whereas the low-performing group's stayed roughly the same across the experiment. Error bars show +/−1 SE.

Experiment 1 and compare them to the scaled error rates in Experiment 2. Fig. 10 displays this cross-experiment comparison and shows a similar pattern across experiments.

We examined errors and correct response times by their sequential position and compared these to RTs from Experiment 1. Overall, there is a significant correlation (r = .88, t(8)=5.37, p < .001) between correct RTs from the RL experiment and RTs from the cued SRT experiment. Comparing the cued RTs to the high- and low-scoring groups separately revealed a difference between the groups. The cued SRT RTs do not correlate significantly with the high-scoring group's RTs (r = .51, t(8) = 1.68, p = .13), but they do correlate significantly with the number of errors made in the RL experiment (r = .83, t(8) =4.18, p < .01). The low-scoring group shows the opposite pattern. The cued SRT RTs correlated significantly with the RL correct RTs (r = .80, t(8) = 3.79, p < .01) but not with the RL errors (r = .57, t(8) = 1.96, p = .09). Comparing the two groups with each other revealed a significant correlation in errors (r = .79, t(8) = 3.68, p < .01), but no significant correlation in RT (r = .17, t(8) = 0.48, p > .05). In conclusion, the RTs in

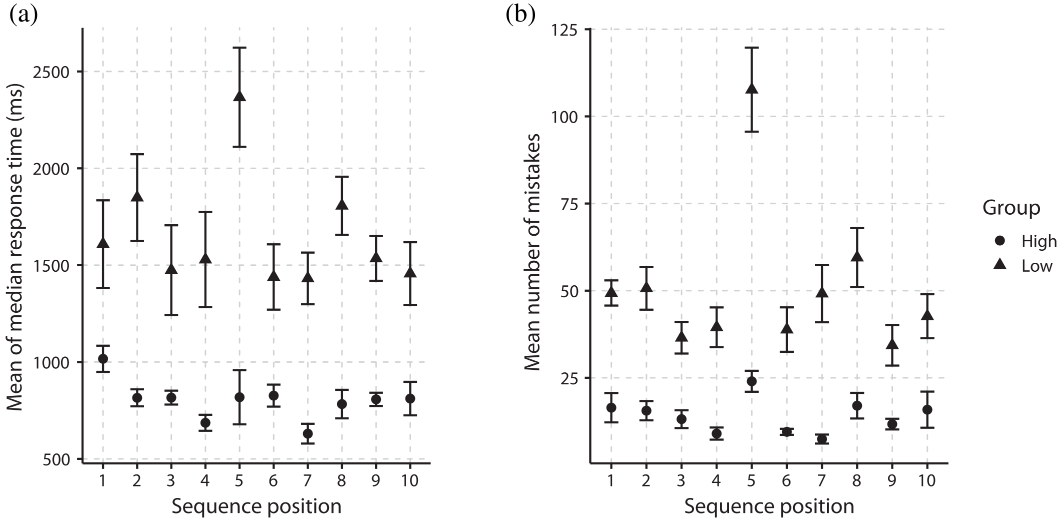

Fig. 9. RTs and error rates by median split and sequential position. (a) Mean of subjects' median correct response times by median split and sequential position. The correct RTs for the two performance groups were not significantly correlated (r = 0.17, t(8)=0.48, p = 0.65). Error bars reflect +/-1 SE. (b) The mean number of errors made at each position in the sequence split by performance group. The errors are highly correlated (r = .79, t(8)=3.68, p < .01), although note how much worse sequence position 5 was for the low-performing group relative to the next-worst position (8). Low-performers showed twice as many errors in position 5 as in 8, while the high-performing group showed only a 25% increase in errors. Error bars reflect +/−1 SE.

Experiment 1 and RTs and error rates in Experiment 2 were correlated, suggesting that the results observed in Experiment 1 are largely due to predictive, rather than reactive processes.

4. Models

4.1. Modeling environment

To compare human sequence acquisition with existing reinforcement learning models, we implemented two reinforcement learning models and a simple negative recency biased model (SCM; Boyer et al., 2005). Fig. 11 illustrates the modeling experiment's setup. The environment contains all data regarding the targets, which it passes to the task, which in turn passes the current state of the environment to the agent, which selects the relevant action. The action is evaluated by the environment, which updates itself and passes a reward to the agent. The reward is used to update the agent's strategy, and the model continues with the next step. We defined the reinforcement learning SRT task in this framework for our simulations.

As in the human experiment, the data regarding the targets was only partially visible to the agent. The task acted as a veil through which a certain state would be observable.

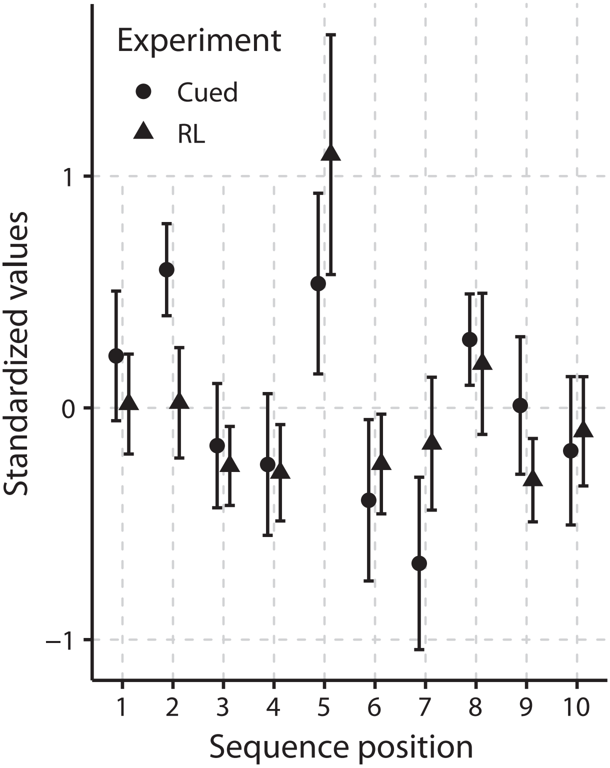

Fig. 10. Scaled mean number of errors in Experiment 2 (RL) against scaled correct RTs from Experiment 1's cued SRT paradigm (NB87) by sequence position. The number of errors per position and the correct RTs are significantly correlated (r = 0.64, t(8) = 2.36, p < .05). Error bars show +/−1*SE*.

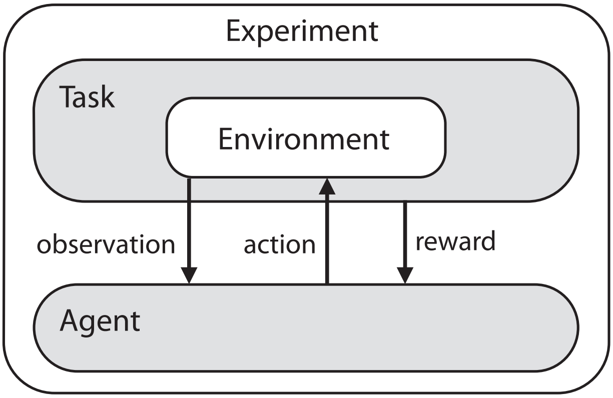

Fig. 11. Overview of the experimental setup for the RL models. The plated components interact with each other according to the arrows to simulate the same trial-and-error learning process that humans undergo.

To a human participant, the current position in the sequence would be obvious, as it was colored differently from the other stimuli. At a minimum, the immediately prior occupied position was probably obvious as well, readily available in memory. Positions preceding that, however, might not be reliably accessible in memory. In the sequence we used (4-2-3-1-3-2-4-3-2-1), following Nissen and Bullemer (1987), each position's identity is fully determined by the previous two positions. That is, one could perfectly predict the next position given only the two prior to it—assuming one has determined that there is a deterministic, periodically repeating sequence. The RL models we use rely on a set of third-order observations, assuming that the models know their current position and the two prior positions. For the RL models, a softmax action selection policy was used for easier comparison with the SCM model.

4.1.1. Simple condensator model

As a baseline model using a simple negative recency bias (avoiding targets that were reached recently), we used a simple condensator model (SCM), introduced by Boyer et al. (2005), and inspired by Dominey (1998). In this model, each target is assigned a corresponding unit, with activation ranging from 0 to 1.0. Summed activation across units is always 1.0, and all units were initialized at 0.25. Each step, the activation across units determines the probability distribution of choosing each target (similar to the softmax action selection policy used in the reinforcement learning models described later). After selecting a target, its corresponding activation is then distributed equally among the other three units, resetting the activation to zero.[3]

4.1.2. On-policy versus off-policy learners

The reinforcement learning models differ in their learning component, which is contained within the agent and maintains a mapping between states and action-values. For each given input-state there are three action-values, corresponding to the number of movements that can be made by the agent. Upon receiving a reward, the agent updates the action-values using its learning algorithm. We tested two learning algorithms: Q-learning (Watkins, 1989) and SARSA (Rummery & Niranjan, 1994).

Off-policy learners such as Q-learning learn the value of the optimal policy independently of the agent's actions. They learn about the greedy policy, updating old action-values using the maximum of all action-values for the current state, while—depending on the action selection policy used—it stochastically selects actions, sometimes exploring. In contrast, on-policy learners (e.g., SARSA) learn the value of the policy actually being carried out by the agent: Instead of the maximum, they take into account the action that was selected for the current state.

Both Q-learning and SARSA are parameterized with learning rate α and discount factor γ. The learning rate determines the weight given to the current information as compared to new information, while the discount factor determines the influence of expected future reward. As with the human participants, the simulated SARSA and Q-learners were tasked with iterating over the repeated sequence until the successful completion of 800 movements. We used a softmax action selection policy, transforming the Q-values for all possible actions in a given state into a probability distribution from which an action was selected.

4.2. Modeling results

To check if our models could account for the wide range of human scores, we optimized parameters by means of a grid search to maximize final score. Scores achieved by Q-learning (*M* = 764, *SD* = 7 with α = .965, γ = .98) were significantly higher than the scores achieved by SARSA (*M* = 753, *SD* = 10 with α = .900, γ = .995), $t(198) = 9.01$, $p < .001$. The maximum score for humans was 725, suggesting that these two models are capable of reaching human performance.

4.2.1. Model fitting procedure

To further investigate the behavior of these models and compare them with human behavior, we fitted the models to individual subjects' behavior during the task. Best-fitting parameters were sought to account for the choice-by-choice actions of each subject. Log-likelihood of each observed sequential choice (correct and incorrect) predicted by the model was calculated, and parameters were sought to maximize the sum, separately for each participant. Subsequently, the Bayesian information criterion (BIC) for each model was calculated as $BIC = -2\ln(\hat{L}) + \ln(n)k$, in which $\hat{L}$ is the maximized likelihood of the model given a participant's observations, n is the number of data points for each participant, and k is the number of free parameters (two for the reinforcement learning models, zero for the SCM model). BIC provides a criterion for model selection using a penalty for the number of free parameters.

4.2.2. Model fitting results

Overall, Q-learning outperformed SARSA for all participants in terms of BIC. Mean optimized learning rate α for Q-learning was 0.323, and mean optimized discount factor γ was 0.958, with a mean BIC of 1920.5 for Q-learning compared to SARSA's mean BIC of 2122.9. For Q-learning, there was a positive correlation between learning rate α and participants' final score on the reinforcement learning task, $r(11) = .703$, $p = .007$, and a negative correlation between discount factor γ and participants' final score, $r(11) = -.574$, $p = .040$, illustrated in Fig. 12. Learning rate α and discount factor γ were not correlated, $p = .323$.

Interestingly, behavior of the four worst-performing participants was best captured by the SCM model, suggesting that these participants employed a simpler strategy by avoiding recent stimuli instead of associating states, actions, and rewards. Detailed information about optimized parameters and model fits can be found in Table A1.

The relationship between learning rate α and discount factor γ and performance on the reinforcement learning tasks shows that best performance requires relatively high values of α and relatively low values of γ. In terms of cognitive mechanisms, this suggests that low-scoring individuals are resistant to updating their action-value function, or the

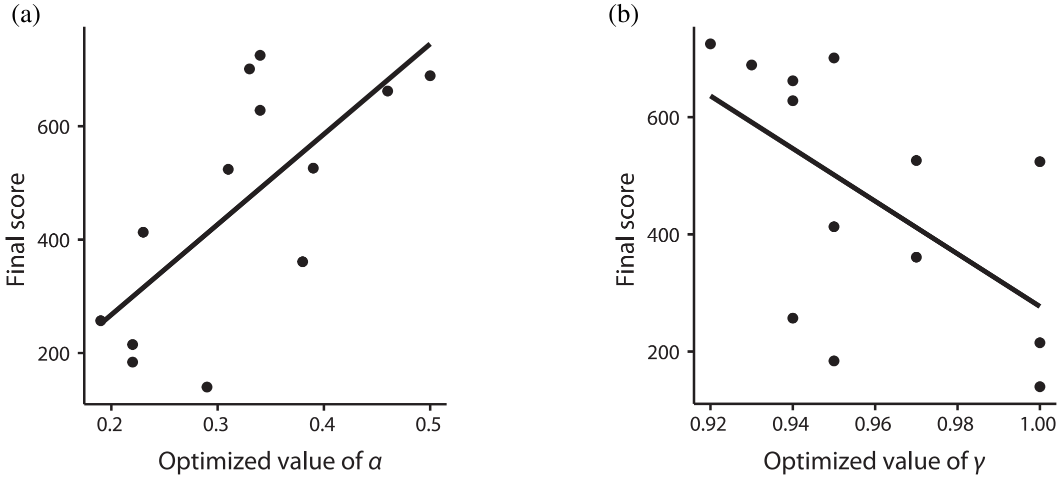

Fig. 12. Relationship between individually optimized parameters for Q-learning and participants' final score. (a) The relation between learning rate α and final score attained by participants. Participants with a higher learning rate α attained higher final scores on the reinforcement learning task. (b) The relation between discount factor γ and final score attained by participants. Participants with a higher discount factor γ attained lower final scores on the reinforcement learning task.

expected utility of their actions. In deterministic tasks such as the current one, an α of 1 would be optimal.

Interesting venues for future research would be finding out to what extent these parameters are affected by task demands versus personal characteristics, for example, by allowing these parameters to drift over time or correlating them with measures of IQ or working memory, such as suggested by Chen (2015).

5. General discussion

Experiment 1 described a trajectory adaptation of the serial reaction time task and found that it replicates the speed-up results of Experiment 1 of Nissen and Bullemer (1987), with participants in the NB87 group showing larger speed-up than the participants in the random group. Also, RTs by sequence position showed a pattern similar to Nissen and Bullemer (1987). Thus, while the trajectory SRT paradigm retains the essence of the original SRT, it also affords the opportunity to measure a variety of more detailed statistics about subjects' continuous motions, similar to, for example, Moisello et al. (2009). Response trajectories can reveal uncertainty, predictive movements, reversals in decision, and other phenomena that may reveal the dynamics of the learning mechanisms at work. The present study examined the average trajectories of two isomorphic vertical movements that appear in the NB87 sequence, as well as in the random condition. The two movements have different subsequent stimuli in the NB87 condition and were thus expected to show a sequential context effect: As participants learn where the next stimulus will be, they may start to move toward this response even as they finish the previous movement—as a piano player may reach for the next key while the current one is being sustained (Soechting, Gordon, & Engel, 1996).

We found not only that the expected context effects had developed by late training, but also evidence of possibly strategic adaptive behavior in the random condition. Many participants in the random condition developed a re-centering approach after each response, waiting for the next (unpredictable) stimulus to appear. In a way, this behavior is optimal, since the center of the screen is as close as possible to all stimuli. As such, we believe this position is influenced by relative frequency, as well as target location. However, future research should shed light on the determinants of this "optimal resting location." Some participants in both conditions showed this behavior to a limited extent early in training, but those trained on the NB87 sequence lost this behavior over time as they learned to predict the location of the subsequent stimulus —hinted at by the decrease in reaction times in this condition and confirmed by the deviation in average trajectory toward the subsequent stimuli. Of the participants in the random condition, two groups could be identified: a centering group and a non-centering group. This might reflect differences in strategy similar to Tubau et al.'s (2007) stimulus-based versus plan-based control mode or Dale et al.'s (2012) reactive versus predictive movements, although it is yet unclear how centering behavior would be classified. How these different behavioral strategies are related could be the focus of future research.

Overall, the behavioral results show a striking similarity to the Nissen and Bullemer (1987) results. The pattern of reaction times over sequence position was almost entirely equivalent to the pattern observed in the original study, although the movement reaction times were higher throughout training and participants showed less overall improvement. This can be explained through the mechanics of the paradigm: Mouse movements require more time to be executed than single keypresses, and they require some fine motor control and error correction. The sensitivity of the mouse can be adjusted to achieve a balance between RT and error; we used a very low sensitivity to reduce overall noise. Participants in the NB87 sequence condition nonetheless showed an increased number of errors during training, indicative of a speed-accuracy trade-off, which was not present in the Nissen and Bullemer (1987) results. It is possible that extending the training would eventually lead to a reduction of errors, as participants would gradually become aware of the sequence.

In Experiment 2, recognizing that everyday action learning is often done by trial and error, we adapted the trajectory SRT paradigm to be a reinforcement learning task. This adaptation allowed us to investigate purely predictive movements, as there was no cue. The task proved to be more challenging for some than for others, as indicated by differences in response times and accuracy. Those data also suggest that participants adopt different strategies and tried to adapt when they were not learning. These findings are similar to those in Experiment 1: RTs in Experiment 1 were correlated with accuracy in Experiment 2. In particular, data from the high-performing participants compared remarkably well to Experiment 1, despite the task differences. The most notable similarity was the difficulty participants experienced with the fifth stimulus position. It could be that the fifth stimulus was on a chunk boundary, or that it was confused with the transition from the last to the first stimulus.

A bimodal distribution of scores showed that half of the participants did really well, as they made very few errors after roughly 10 repetitions of the sequence. Block-by-block analysis of the response times showed a difference in speed-up across the experiment, indicating the high-performing group learned the sequence much better than the low-performing group. The difference in response times to incorrect targets suggests the two groups might have used different strategies. The rare but increasingly slow errors in the high-performing group suggest more time was spent figuring out the next stimulus, while the persistent and relatively fast errors of the low-performing group suggest participants may have adopted a probabilistic view of the task, randomly trying options instead of trying to learn a deterministic pattern.

Despite the major difference of no cueing of the next response, performance in the RL experiment was quite comparable to performance in the cued SRT experiment. This suggests that a large component of implicit learning in tasks like this is prediction-related, and not merely a result of pre-potentiation. If this is indeed the case, reinforcement learning tasks can be used in sequential paradigms in order to investigate purely predictive actions. The pattern of correlations indicated a difference between the low- and high-performing groups that was not immediately obvious. Overall, the cued SRT response times are correlated to RTs and accuracy data from the RL experiment, whereas this is not true for both the low- and high-performing groups separately. We expect this is due to different strategies among groups, leading to a different pattern of speed and accuracy at different sequence positions.

In addition to our behavioral analyses, we tested three different models to see if human behavior could be explained by simple, model-free responses to sequential stimuli. The two reinforcement learning models, Q-learning and SARSA, performed quite well, with Q-learning outperforming SARSA both in maximum score as well as in model fits. The results suggest that sequence learning in the current paradigm can be partially explained by model-free reinforcement learning models. Behavior of the four worst-performing participants was better captured by a condensator model with negative recency bias, suggesting that instead of learning the sequence by state-action associations, poor performers used simple heuristics in the reinforcement learning task. Future studies could shed light on the role of goals in the acquisition of such action sequences and the way learning shifts from simple to more complex mechanisms, as has been shown to exist for single-step action (see, e.g., Hommel, Müsseler, Aschersleben and Prinz (2001) for one proposed mechanism of goal-directed action). Also, model-based reinforcement learning models may be able to account for additional, explicit, learning. The process by which humans acquire action sequences seems to vary, and the diverse strategies employed evidently yield quite variable performance. However, studying sequence learning is important, as most of human behavior—and experience—is quintessentially sequential.

Acknowledgments

The preparation of this work was supported by the European Commission (EU Cognitive Systems project ROBOHOW.COG; FP7-ICT-2011). Experiment 1 was presented previously in a conference paper (Kachergis et al., 2014b), but it appears here with additional analyses.

Notes

1. There was 33% of chance success in one try (+1), 33% chance of success in two tries (−1+1), and 33% chance of success in three tries (−1−1+1).
2. Hartigan's dip test did not show final scores to significantly differ from a unimodal distribution; however, our limited sample size limits the usefulness of this test.
3. More specifically, the activation was reset to a small constant to avoid division-by-zero errors in calculating the log-likelihood during model fitting.

References

Averbeck, B. B., & Costa, V. D. (2017). Motivational neural circuits underlying reinforcement learning. *Nature Neuroscience*, *20*, 505–512.

Bornstein, A. M., & Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience*, *35*, 1011–1023.

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to routine sequential action and its pathologies. *Psychological Review*, *111*, 395–429.

Boyer, M., Destrebecqz, A., & Cleeremans, A. (2005). Processing abstract sequence structure: Learning without knowing, or knowing without learning? *Psychological Research, 69,* 383–398.

Bruhn, P., Huette, S., & Spivey, M. (2014). Degree of certainty modulates anticipatory processes in real time. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 525–538.

Chen, C. (2015). Intelligence moderates reinforcement learning: A mini-review of the neural evidence. *Journal of Neurophysiology*, *113*, 3459–3461.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253.

Cooper, R. P., & Shallice, T. (2000). Contention scheduling and the control contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*, 2987–338.

Dale, R., Duran, N. D., & Morehead, J. R. (2012). Prediction during statistical learning, and implications for the implicit/explicit divide. *Advances in Cognitive Psychology*, *8*, 196–209.

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences, and reinforcement learning. *European Journal of Neuroscience*, *35*, 1036–1051.

Dominey, P. F. (1998). Influences of temporal organization on sequence learning and transfer: Comments on Stadler (1995) and Curran and Keele (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 234–248.

Duran, N. D., & Dale, R. (2009). Anticipatory arm placement in the statistical learning of position sequences. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 893–898). Amsterdam: Cognitive Science Society.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*, 447–455.

Fendrick, P. (1937). Hierarchical skills in typewriting. *Journal of Educational Psychology*, *28*, 609–620.

Fischer, M. H., & Hartmann, M. (2014). Pushing forward in embodied cognition: May we mouse the mathematical mind? *Frontiers in Psychology, 5,* 1315.

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex*, *22*, 509–526.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*, 614–636.

Fu, Q., Fu, X., & Dienes, Z. (2008). Implicit sequence learning and conscious awareness. *Consciousness and Cognition*, *17*, 185–202.

Gehring, W. J. (1992). The error-related negativity: Evidence for a neural mechanism for error-related processing (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Gentner, D. R., LaRochelle, S., & Grudin, J. (1988). Lexical, sublexical, and peripheral effects in skilled typewriting. *Cognitive Psychology*, *20*, 524–548.

Gureckis, T. M., & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*, 293–313.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679–709.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*, 849–937.

Kachergis, G., Berends, F., de Kleijn, R., & Hommel, B. (2014a). Reward effects on sequential action learning in a trajectory serial reaction time task. *IEEE Conference on Development and Learning / EpiRob 2014*.

Kachergis, G., Berends, F., de Kleijn, R., & Hommel, B. (2014b). Trajectory effects in a novel serial reaction time task. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 713–718.

Kachergis, G., Wyatt, D., O'Reilly, R. C., de Kleijn, R., & Hommel, B. (2014). A continuous time neural model for sequential action. *Philosophical Transactions of the Royal Society B*, *369*, 20130623.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley.

Marcus, D. J., Karatekin, C., & Markiewicz, S. (2006). Oculomotor evidence of sequence learning on the serial reaction time task. *Memory & Cognition*, *34*, 420–432.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533.

Moisello, C., Crupi, D., Tunik, E., Quartarone, A., Bove, M., Tononi, G., & Ghilardi, M. F. (2009). The serial reaction time task revisited: a study on motor sequence learning with an arm-reaching task. *Experimental Brain Research*, *194*, 143–155.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. *Cognitive Psychology*, *19*, 1–32.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–263.

Robinet, V., Lemaire, B., & Gordon, M. B. (2011). MDLChunker: A MDL-based cognitive model of inductive learning. *Cognitive Science*, *35*, 1352–1389.

Rummery, G. A., & Niranjan, M. (1994). On-line Q-Learning using connectionist systems (Tech. Rep. No. CUED/F-INFENG/TR 166). Cambridge University.

Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*, 163–196.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489.

Skinner, B. F. (1950). Are theories of learning necessary?. *Psychological Review, 57,* 193–216.

Soechting, J. F., Gordon, A. M., & Engel, K. C. (1996). Sequential hand and finger movements: Typing and piano playing. In J. R. Bloedel, T. J. Ebner, & S. P. Wise (Eds.), *The acquisition of motor behavior in vertebrates* (pp. 343–360). Cambridge, MA: MIT Press.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, *15*, 207–211.

Stadler, M. A. (1992). Statistical structure and implicit serial learning. *Journal of Experimental Psychology*, *18*, 318–327.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tubau, E., Hommel, B., & López-Moliner, J. (2007). Modes of executive control in sequence learning: From stimulus-based to plan-based control. *Journal of Experimental Psychology: General*, *136*, 43–63.

Watkins, C. J. C. H. (1989). Learning from delayed rewards (Unpublished doctoral dissertation). Cambridge University.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1047–1060.

Appendix A: Optimized parameters and model fits for all participants

Table A1
Optimized parameters and Bayesian information criterion (BIC) for each participant and model

		Q-learning			SARSA			SCM
Participant	Score	α	γ	BIC	α	γ	BIC	BIC
1	524	0.31	1.00	1671.9	0.47	1.00	1901.9	1992.0
2	140	0.29	1.00	3596.5	0.33	1.00	3914.5	2534.0
3	689	0.50	0.93	857.6	0.55	0.94	963.6	1622.0
4	215	0.22	1.00	3240.4	0.34	1.00	3564.4	2740.0
5	257	0.19	0.94	3162.4	0.34	0.95	3526.4	2570.0
6	628	0.34	0.94	1229.7	0.44	0.97	1395.7	1810.0
7	184	0.22	0.95	3358.5	0.33	0.99	3636.5	2666.0
8	662	0.46	0.94	1011.6	0.53	0.95	1133.6	1732.0
9	725	0.34	0.92	727.5	0.48	0.96	813.5	1552.0
10	526	0.39	0.97	1731.9	0.54	0.99	1903.9	2086.0
11	413	0.23	0.95	2340.1	0.29	0.99	2566.1	2330.0
12	361	0.38	0.97	1213.7	0.39	0.98	1365.7	1832.0
13	701	0.33	0.95	823.6	0.39	0.99	911.6	1622.0