Modelling the Decay of Declarative Knowledge Between Learning Sessions

Felix Boie

s2879409

29 December 2020

Master Thesis

Human-Machine Communication

University of Groningen, The Netherlands

Internal supervisor:

Dr. Jelmer Borst (Artificial Intelligence, University of Groningen)

External supervisors:

Maarten van der Velde, MSc. (Experimental Psychology, University of Groningen)

Prof. dr. Hedderik van Rijn (Experimental Psychology, University of Groningen)

Abstract

Adaptive learning systems adapt to the individual learner to optimize learning outcomes. By modelling the strength of facts in human memory over time, an adaptive system can present facts at the optimal time, right before they are forgotten. Prior studies have shown that facts decay more slowly between learning sessions than within learning sessions. In this thesis, we confirm this finding in naturalistic learning data collected using an adaptive learning system in two university courses. We demonstrate that while the system's current ACT-R memory model captures within-session memory performance well, it does not adequately capture between-session memory decay. To better account for this, we extend the model by scaling the time between sessions by a psychological time factor (PTF). Here we show in detail how the PTF improves the learning system. Our findings suggest that the optimal PTF depends on the interval between sessions and is affected by sleep. Specifically, the PTF decreases as sessions are spaced further apart and remains constant once learners are thought to have slept between sessions. This thesis demonstrates the need of accounting for the passage of time in a more refined way than just using a single scaling factor for the PTF.

Modelling the Decay of Declarative Knowledge Between Learning Sessions

We often need to learn new facts: birthdays of friends and family, street names of a city, vocabulary for a new language and more. Learning these facts can be tedious and sometimes quite time consuming. Preferably, you could optimize the time you spend on studying, meaning study in a way that does not take much time, while still allowing you to remember the facts for a long time. So how can you learn facts well and efficiently?

Robust memory effects that can help with improving learning strategies are the spacing and testing effect (Delaney, Verkoeijen, & Spirgel, 2010). The spacing effect states that spacing out learning encounters over time compared to massing them helps people to remember these facts over a longer time period. The testing effect states that active retrieval of facts is more effective for long term retention compared to passive studying (Karpicke & Roediger, 2008; Roediger & Butler, 2011). This implies for learning that you should not just read over your learning material, but try to actively recall it to get the best learning effects. The testing effect is strongest if the learner can still successfully recall an item (Carrier & Pashler, 1992) and disappears if the learner is unable to recall the item (Jang, Wixted, Pecher, Zeelenberg, & Huber, 2012). This suggests that for the best learning outcome you should space out the repetitions of facts as far as possible over time (spacing effect), but only up to the point, where the learner can still recall them (testing effect).

Model-based learning systems can make use of these effects to optimize learning. These systems use computational models of learning and forgetting in human memory to help with the learning process. These systems might assume that across different learners and materials forgetting and learning follow qualitative similar patterns, but that these patterns differ in their quantitative properties between learners and materials. A system might for example assume that declarative knowledge is forgotten in the form of a power curve (Anderson & Schooler, 1991) though the precise decay might differ between participants and items. Using these models the system can then present items to the learner for which it expects the best learning outcome. For example the system MemReflex (Edge, Fitchett, Whitney, & Landay, 2012) optimizes the

learning for microlearing (very short learning sessions). The system estimates how likely the learner will be able to recall an item at a specific time an tries to keep this likelihood at 90%. The idea is that the learner is kept motivated and ideally in a state of flow.

Another system, which has been shown to help learners with studying is an adaptive fact learning system developed at the University of Groningen (Sense, Behrens, Meijer, & van Rijn, 2016; Van Rijn, van Maanen, & van Woudenberg, 2009). This system is the main focus of this thesis and will be referred to as the RUGged learning system. The RUGged learning system is an extension of the adaptive item-learning model by Pavlik and Anderson (2005, 2008). It has successfully been used in laboratory (Sense et al., 2016; Sense, Meijer, & van Rijn, 2018; Van Rijn et al., 2009) and field (Sense, van der Velde, & van Rijn, 2018) research. In the RUGged learning system an activation function represents the availability of facts in memory of the learner over time. Based on these activations the system tries to present facts right before they are forgotten, which should lead to the best learning outcomes.

Learning systems typically focus on one of two approaches for distributing learning material over time: optimize spacing of facts within sessions (Donovan & Radosevich, 1999) or optimizing the distribution of sessions (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). A learning session describes the time period in which a learner focuses on studying a given learning material. While this can be defined quite clearly in a laboratory setting (for example the participant should study a list of words from time x till time y), it is more ambiguous in a less controlled setting. The RUGged learning system is optimized for within session spacing (e.g., Van Rijn et al., 2009). However, in practice learning is often spread out over multiple sessions. While the framework works well within the time span of seconds to minutes (Van Woudenberg, 2013), it has problems to accurately model the activation of facts in the learner's memory over longer periods of time (Anderson, Fincham, & Douglass, 1999).

One problem is that the model predicts a quicker decay of activation between sessions, than what seems to happen in reality (Anderson et al., 1999). This means that the model predicts that a learner forgets a fact quicker than he or she actually does.

This can lead to potential problems: items are not optimally spaced (so they are repeated too early) and learners might get annoyed by needing to restudy an item, which they still know quite well.

Anderson et al. (1999) and Pavlik and Anderson (2003) suggested the construct of psychological time to account for this slower memory decay for the interval between sessions. They slowed the decay between sessions by scaling the time between sessions for the model by a factor smaller than 1, therefore reducing the time between sessions from the model's perspective. The underlying idea of psychological time is, that during a learning session you encounter more interference from intervening memory events compared to outside the learning sessions. The scaling factor should account for this different interference by representing a value of how much the time "chips away" at the memory traces compared to the time during the learning session. As this scaling factor is only applied to the time between sessions, it will only change the predictions of the model for items after the first session they were encountered. Using this method Anderson et al. (1999) and Pavlik and Anderson (2003) were able to better model the probability of their uses to answer the provided items correct or incorrect over multiple sessions.

In this thesis, we investigated the effects of expanding the RUGged learning system with the construct of psychological time. First, we will outline the RUGged learning system and the concept of psychological time in more detail. We will then describe the effects of expanding the RUGged learning system with psychological time on two large datasets, which were collected in realistic settings. Afterwards, we will present a way of fitting psychological time to individual learning sessions and investigate factors that influence the fitted psychological time. Finally, the results will be summarized and viewed in a greater context of memory research.

RUGged learning system

The RUGged learning system makes use of the spacing and testing effect to optimize the introduction and repetitions of facts. The system is based upon the

spacing system of Pavlik and Anderson (2008) and both are based on the declarative memory equations of the cognitive architecture ACT-R (Adaptive Character of Thought - Rational) (Anderson, 2007). In the ACT-R architecture each memory trace for an item receives an activation value over time. The higher the activation value the more likely it is that the participant will be able to recall the item correctly at that time. An item has the highest activation at the moment it is encountered. Afterwards, it decays over time according to the following equation:

$$A_i(t) = ln(\sum_{j=1}^{n}(t - t_j)^{-d_j}) \tag{1}$$

The activation A of item i at time point t is dependent on the prior encounters of the same item (n = number prior encounters, t_j = time of prior encounter, d_j = decay of prior encounter). The decay of a prior encounter is evaluated at the time it was encountered based on the equation:

$$d_j = c \cdot e^{A_i(t)} + \alpha \tag{2}$$

The Equation 2 for decay consists of an activation-dependent component and a separate item-specific offset (α) that captures the rate of forgetting. The parameter c is fixed and represents the decay scale parameter, which determines the relative contribution of the activation-dependent component. In Equation 2 the activation (A) is calculated based on the Equation 1 over all previous encounters of item i, excluding the current encounter for which the decay is evaluated. The Equation 2 shows that the decay of an encounter depends on how active the item is in memory at that point in time. If an item has a high activation when it is encountered, then this encounter will receive a higher decay rate. In other words, when an item has a high activation when it is encountered, then it will improve the retention of the item less compared to an encounter, where the item had a lower activation. This way the spacing effect is accounted for. Spacing out items over longer periods of time makes sure that the activations of items are lower when they are encountered, which consequently leads to lower decay rates.

The item-specific offset α is by default 0.3 and is adapted based on the responses of the learner, specifically the response time of the learner and whether the item was answered correct or not. The RUGged learning system estimates the response time based on the equation for retrieval time from ACT-R and an added fixed time (Equation 3). The retrieval time depends on the activation A of an item i at time t and is scaled by the parameter F. If the activation is high for an item, it will be recalled faster. A fixed time is added to account for the time of non-memory related tasks: processing the visual stimulus and pressing a key (Nijboer, 2011). Pavlik and Anderson (2003, 2005, 2008) showed that these equations can account for a wide range of data for learning related experiments.

$$L_i(t) = F \cdot e^{-A_i(t)} + \text{ fixed time} \tag{3}$$

Before addressing latency, we will first focus on accuracy. Based on these equations Pavlik and Anderson (2008) created an adaptive fact-learning system, which can be used to create an optimal schedule of presenting facts to the learner. Their system adjusted the decay parameter based on the accuracy scores of the learners. Their system simulates the activation of items over time and uses these estimations to predict if the learner will get an item correct or incorrect. If an incorrect response is given to an item for which the system would have predicted that the learner should answer it correctly, then the simulated activation is too high. The estimation would have fitted the given response better if the decay of this item would have been higher. The activation is lowered if the learner answers an item correctly for which the system had predicted that the learner could not recall it anymore. See Pavlik and Anderson (2008) for a full explanation of the system.

The RUGged learning system extends the system of Pavlik and Anderson (2008) by also taking the response time into consideration. Based on Equation 3 the system estimates how quickly the learner will respond at an encounter. This expectation can then be compared to the actual response time of the learner. The RUGged learning system uses a binary search algorithm to adjust the α parameter to fit its predicted

response time to the observed response time (Nijboer, 2011). For example, if an item is answered correctly but slower than predicted by the RUGged learning system, the α parameter will be increased. This allows a more fine-tuned adjustment for correct responses compared to the prior method of Pavlik and Anderson (2008). This is important because learners usually provide more correct than incorrect responses.

Psychological time

Usually effects of within-session trial distribution (Donovan & Radosevich, 1999) have been considered independently of effects of between-session practice (Cepeda et al., 2006). Attempts to combine both time scales into one model, seem to suggest that the rate at which a learner forgets items slows down after some time (Elliott & Anderson, 1995; McBride & Dosher, 1997). This is a problem when using the ACT-R equations over longer time periods because ACT-R assumes the same rate of forgetting over different time scales. While these equations model learners' behavior relatively well within shorter intervals (range from seconds to minutes), they seem to work less well for longer time periods. Its simulation of decay of memory traces between sessions would lead to too low activations compared to what has been found in studies (Anderson et al., 1999; Pavlik & Anderson, 2003). This can be seen in that the response times of participants after a long interval between sessions are usually quicker than predicted by the ACT-R equations. This can be problematic for the RUGged learning system because it relies on these equations to space the repetition of facts optimally. In practice this could lead to an inefficient allocation of study time and frustration of learners because they are asked to restudy facts they still know well enough.

To account for this slower decay between sessions Anderson et al. (1999) proposed the construct of psychological time and developed the idea in Pavlik and Anderson (2003) further. The idea is that forgetting is not only based on the passage of time, but also upon the interference of memory intervening events within this time. The proposed psychological time factor (PTF) represents how quickly intervening events "chip away" at the memory traces compared to during the learning session. The effect of interference

seems to be stronger for more similarity between learned material and intervening events (Rasch & Born, 2013). As the events during a learning session will be more similar to the learned material compared to outside the session, one can assume that there is also less interference outside a learning session. They made the distinction for the time within a session, which passes normally and the time between sessions, which is scaled by a PTF. The PTF should be between 0 and 1. A PTF of 0.01 would mean that the interfering events are 100 times less influential between sessions than within sessions. This is modeled by multiplying the interval between sessions with the PTF: for example 100 minutes of actual time with a PTF of 0.01 becomes one minute for the model. In similar kind of learning studies, PTFs that have been used were 0.00046 (Pavlik, Bolster, Wu, Koedinger, & Macwhinney, 2008), 0.0172 (Pavlik & Anderson, 2008), 0.025 (Pavlik & Anderson, 2005) and 0.031 (Pavlik & Anderson, 2003). It is important to highlight that the order of magnitude of the different PTFs differs quite a bit between experiments, indicating that there might not be one PTF fitting in every circumstance. These PTFs were used to improve the model's predictions to the responses of the learners within a session. The interval between sessions in these experiments was one day or more. Figure 1 shows how the activation of a single encounter of an item would decay over time for different α values and different PTFs. Lower PTFs lead to a lower decay of activation over time. Furthermore, for lower PTFs the difference between the selected α values is also smaller.

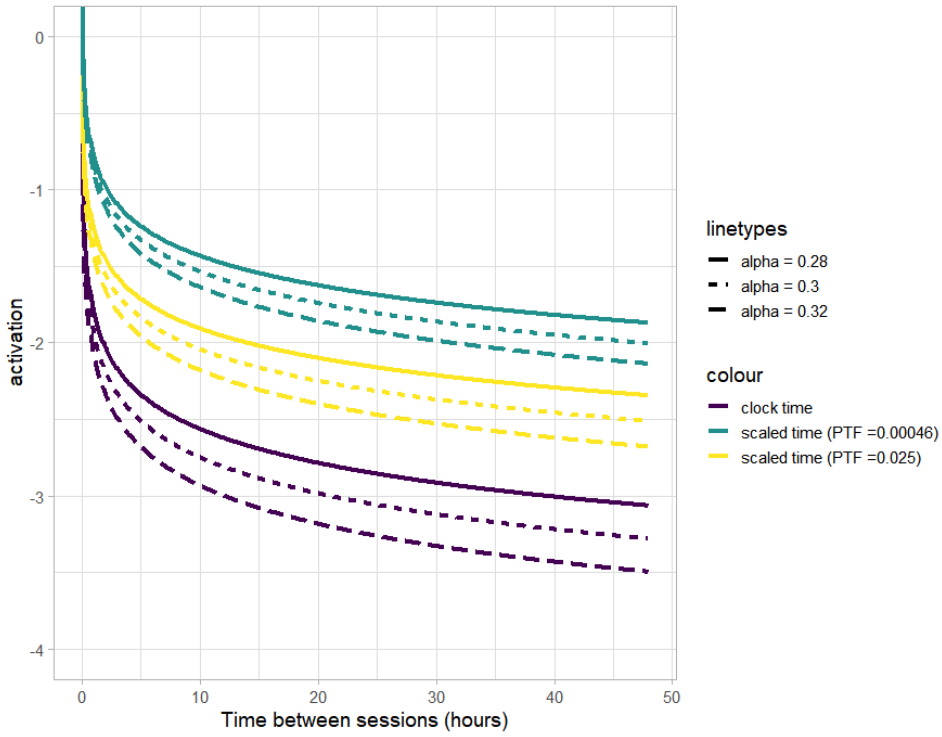

Figure 1. This graph shows the effect of applying psychological time on the decay of activation over time. The activation is based on a single encounter at t = 0. The colors show different PTFs (clock time: PTF = 1). The different line types show the result if we assume different α values.

Applying psychological time to the RUGged learning system

Methods

With psychological time the whole study time for the learner consists of the time within learning sessions and the time between learning sessions scaled by the PTF. The PTF was added to the RUGged learning system by changing the presentation time (t) of encounters. The presentation time keeps track of when items were presented to the learner. The presentation time scaled with psychological time will be called t_s. The update was applied to whole sessions and was based upon the interval or intervals between sessions before the current session according to following equation:

$$t_s = t - (interval_{\text{between sessions}} - PTF \cdot interval_{\text{between sessions}}) \quad (4)$$

Figure 2 shows the effect of applying a PTF of 0.01 to two sessions, which have an interval between sessions of 100 minutes. In the top condition no scaling was applied and in the bottom condition a PTF of 0.01 was applied. The scaling did not influence the first session but subtracted $100 - 0.01 \cdot 100 = 99$ minutes from the presentation times of every subsequent encounter.

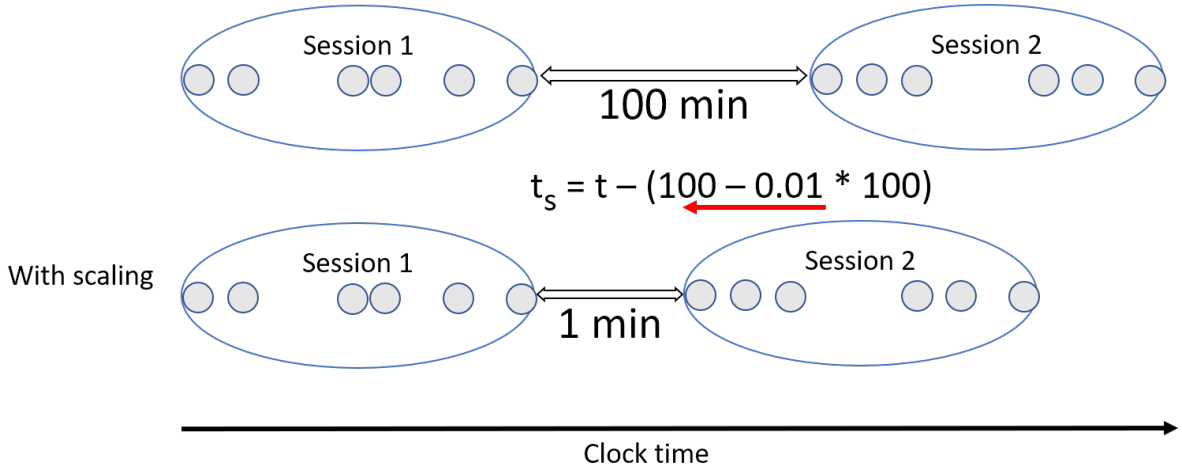

Figure 2. This Figure shows the effect of applying a PTF of 0.01 to two sessions with an interval between sessions of 100 minutes. The grey circles show presentation times of individual encounters in the sessions. The red error indicates the adjustment applied to the presentation time of every encounter in the second session.

Datasets. We investigated the effect of PTF on the RUGged learning system using two prior collected datasets. In these datasets students from the University of Groningen studied with the RUGged learning system in realistic conditions. Students decided themselves when and how long they wanted to use the system. Only data of students who provided informed consent was analyzed.

The first dataset (cogPsych) contains the data of 154 students studying with the RUGged learning system for a cognitive psychology exam in the years 2017 and 2018. Students were free to use the system as they saw fit and could access it online. The provided learning material was grouped by chapters of the textbook for the course and introduced after that chapter was discussed in class. In total 468,617 learning encounters were recorded. For more information see Sense, van der Velde, and van Rijn

(2018).

The second dataset (bioPsych) contains the data of 322 students studying with the RUGged learning system course material of the Biopsychology course at the University of Groningen in the year 2019/2020. The material was grouped into chapters from the coursebook. Students needed to achieve two mastery credits per chapter to be allowed to take the Biopsychology exam. Mastery credits were awarded for studying with the RUGged learning system based on an accuracy-based (responding correctly to the last three presentations of all items) and an effort-based (studying a chapter long enough) criterion. The studied material was not directly assessed in the exam. Each student could reach a maximum of one mastery credit per chapter in 24 hours. This criterion rewarded a repeated study of the same chapter over multiple days. This made it more likely to have intervals of varying lengths between learning sessions, which allowed us to investigate the effect of applying psychological time to a variety of intervals between sessions. In total 853,805 learning encounters were recorded in the bioPsych dataset.

Preprocessing. Table 1 shows how many trials are left after each preprocessing step. Both datasets contained chapters with either multiple choice or open response questions. In open response questions the student needed to type in the answer. In multiple choice questions the student received four different answers and needed to select one. Multiple choice questions could allow the student to recognize a response rather than actually retrieve it from memory. While this is not necessary a problem, it might introduce an extra layer of variance in the data. For this reason we removed all items with multiple choice responses from further analysis. This removed 91,568 trials (19.5%) in the cogPsych dataset and 410,938 trials (48.1%) in the bioPsych dataset. In this section the percentage of trails removed by each preprocessing step was based on the remaining trials and not all trials.

The PTF is applied to the interval between sessions. The datasets do not provide a direct indication of the time between sessions because they have no indication of when a session ended for the student. While from the perspective of the students there was a clear start and end of a learning session, the responses were logged individually on a

central server without such an identification. To reconstruct the boundaries between the sessions, the following criteria were used: The end of a session was reached if the presentation time from one encounter to the next encounter in the same chapter for a student was longer than 15 minutes or if another chapter was studied by the student within this time. The interval between sessions was the interval from the last presentation time on the first session to the first presentation time the next time this session was repeated by this learner. To investigate the effect of psychological time a student should have studied a chapter at least twice. Furthermore, the sessions should have a minimal duration to make sure that the student actually studied. For this reason only chapters of students were kept that were studied at least twice for a duration of at least five minutes. This criterion removed 91,952 (24.4%) trials from the cogPsych dataset and 27,399 (6.2%) trials from the bioPsych dataset.

Table 1

Trials left after each preprocessing step

Step	cogPsych	bioPsych
All	468,617	853,805
Not multiple choice	377,049	442,867
Studied twice 15+ minutes	285,097	415,468
Two repetitions of a chapter	180,576	261,228
>250ms	170,866	239,392
<15s	151,121	221,370
Not NA	141,631	211,991
Correct responses	122,938	201,674

The main focus of this thesis was on the first two repetitions of a chapter. These allowed for a more straightforward interpretation of psychological time. After the second repetition of a chapter there are multiple between-session intervals and for each a PTF needs to be found. These PTF then also influences subsequent PTF. Looking only at the first two repetitions of a chapter allowed us to focus on fitting only one PTF

without any interactions between PTFs. Considering only the first two sessions of a chapter removed further 104,521 (36.7%) trials in the cogPsych dataset and 154,240 (37.1%) trials in the bioPsych dataset.

While the RUGged learning system was applied to the whole dataset, only encounters with a reasonable response time were considered in the following analysis. Impossibly fast responses (<250 milliseconds; cogPsych 9710 [5.4%] and bioPsych 21836 [8.4%] of trials) and very slow responses (>15 seconds; cogPsych 19745 [10.9%] and bioPsych 18022 [6.9%] of trials) were removed because they might have indicated non-genuine on-task behavior. Also trials for which no response times were recorded were removed (cogPsych 9490 [5.3%] and bioPsych 9379 [3.4%] of trials).

Furthermore, only correct responses of learners were used because for incorrect encounters the observed response times did not necessarily reflect the activation of the memory traces of the items in question. This removed further 13.2% of trials in the cogPsych and 4.9% in the bioPsych dataset. This left 122,938 trials for the cogPsych dataset and 201,674 trials for the bioPsych dataset.

A fact in the RUGged learning system can theoretically have a negative decay value, which would cause counter-intuitive results with scaling the time between learning sessions. A negative decay value would suggest that the activation of an item is increasing over time. Reducing the PTF for this item would increase the predicted response time instead of decreasing it, which is the effect the PTF should have. The minimal α value was set to 0 to avoid this problem. In the cogPsych dataset this did not have an influence and in the bioPsych dataset 8,552 (4.0%) of trials had an α of 0 if no psychological time was included.

Results

An exemplary PTF was used to investigate the effects of psychological time on the cogPsych and bioPsych datasets. Does this improve the fit between the observed and predicted response times of the RUGged learning system? And furthermore how does the inclusion of psychological time affect the predicted response times in detail? To

show the effects on the fit between predicted response times and observed response times a model with no psychological time (PTF = 1) (clock time condition) and a model with psychological time (PTF = 0.01) (psychological time condition) were compared. The value of 0.01 was chosen because it falls in the range of prior used PTFs (0.00046 - 0.031) and was intended to provide an idea of the effect of psychological time. The expectation was that a model without psychological time predicts too long response times compared to the observed response times because it did not account for the slower decay of activation between sessions. The next section focuses on fitting a PTF.

Figure 3 shows the difference in predicted response time of the RUGged learning system compared to the actual response time of the learner. A value of 0 indicates that the predicted response time fits the observed response time of the student perfectly, while a higher value indicates that the predicted response time is too long. Figure 3 shows the results for the first session and for a different number of presentations of items in the second session. In the first session psychological time did not play a role. The first session shows the results after the third presentation of items because the system adjusts its α only after a fact has been seen three times. Only results for correct responses of students were used because the response time of incorrect responses might not reflect the activation of current items in learner's memory. Additionally, results of the second session only show items which have been seen by the learner in the first session already, otherwise the PTF would not have an effect on these items.

Figure 3 shows that in the clock time condition the observed response times of the first presentation in the second session tended to be heavily overpredicted. In the psychological time condition this strong overprediction was gone for the first presentation. For the first presentation there was also the biggest difference between the clock time and the psychological time condition (difference of medians in bioPsych 2.69 and in cogPsych 5.02). After the first presentation both the clock and psychological time condition predicted the observed response roughly as well as in the first session. It could be argued that there was still some overprediction for the second presentation, though the overprediction was way weaker compared to the first presentation. This

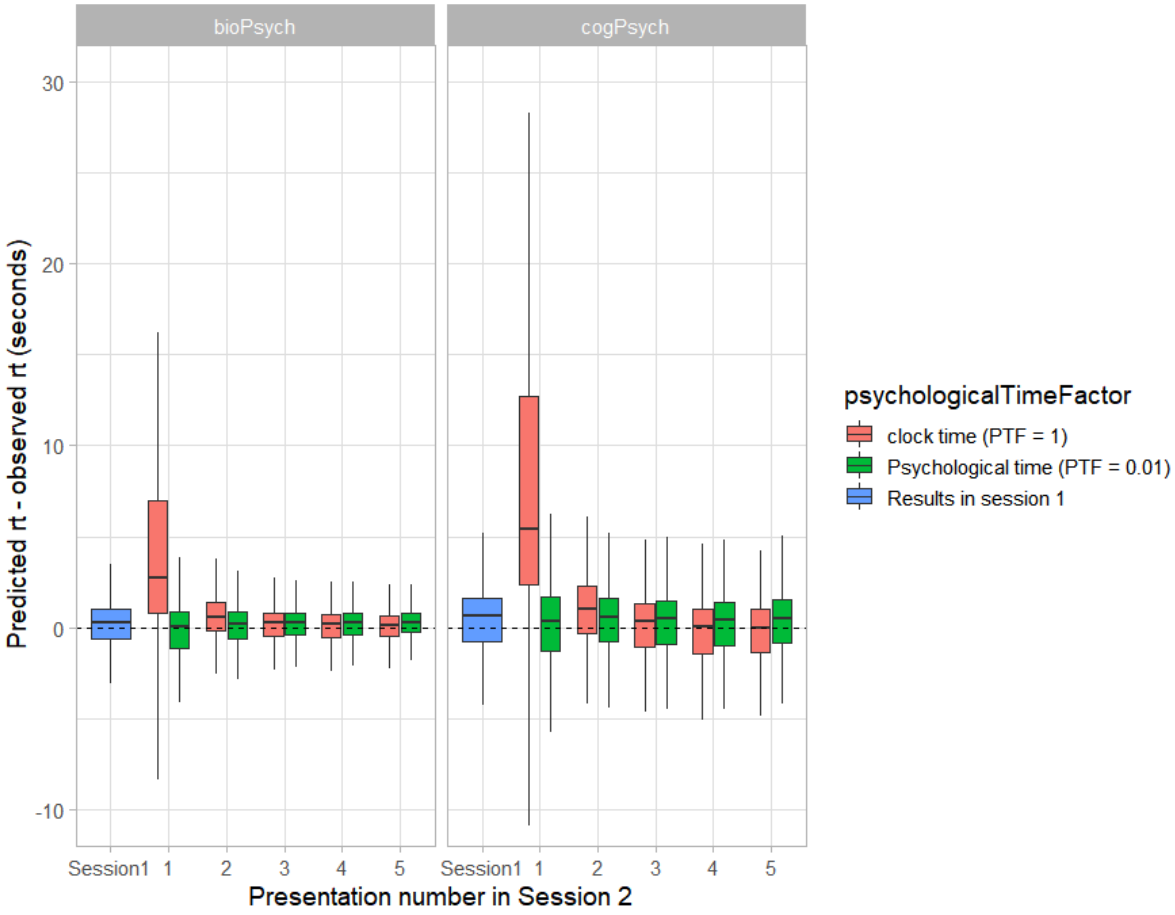

Figure 3. The difference between predicted response time by the RUGged learning system and observed response time of students (higher values indicate that the response time was overestimated by the RUGged learning system) is plotted over the presentation number of items. The left graph shows the result for the bioPsych dataset and the right sigh for the cogPsych dataset. In blue, the difference for the first session is shown. In red (clock time) and green (PTF = 0.01) the difference, aggregated by the repetition of the fact in the second session, is shown. The whiskers of the boxplots are drawn based upon the largest observed point, which falls within the range of $1.5 \cdot IQR$ above the 75% quantile and the lowest observed point, which falls within the range of $1.5 \cdot IQR$ below the 25% quantile. Points outside the whiskers are removed.

pattern was the same in both datasets, indicating that psychological time improved the predicted response times most for the first presentation of items in a session after an interval. The PTF was less influential for higher repetitions because their activation

value was strongly influenced by the prior encounters within this session and because the RUGged learning system adjusted itself to the difference between observed and predicted response times.

The RUGged learning system adapts to the students' responses through its update function of the α value. The α parameter is adjusted to the misfit between predicted and actual response time after every encounter. Figure 3 shows that without psychological time (clock time) this difference tended to be higher. Therefore the system adjusted its α parameter generally more for the clock time condition, than for the psychological time condition. This can be seen in Figure 4. Here the change of α after encounters was plotted for encounters in the first session after the third encounter and for different presentations in the second session. In the psychological time condition the median of the change of the α value was close to zero indicating that in this condition the α value was less adjusted. On the other hand, in the clock time condition the medium value for the first presentation was at -0.05, which is the maximal adjustment for each encounter. This indicates that the α was strongly adjusted for the first presentation. Even for the second presentation the median of the α changed was not at the zero point indicating that the system still needed to adjust its α parameter for most encounters. Afterwards, the change in α became comparable to the first session and the psychological time condition. This indicates how the RUGged learning system currently deals with the slower decay of activation between sessions: it adjusts its α value very rapidly to adjust for the overprediction of response times. With a PTF the α parameter would not need to account for the differing decay of activation between sessions compared to within sessions, allowing a more fine tuned adjustment to the individual students and items.

It is clear that psychological time can improve the predicted response times of the RUGged learning system. The improvement was mainly due to the improved fit of the first presentation of an item after an interval. To get a more in-depth understanding of psychological time, we focused on the effect of psychological time over different intervals between sessions and different observed response times. For this only the first

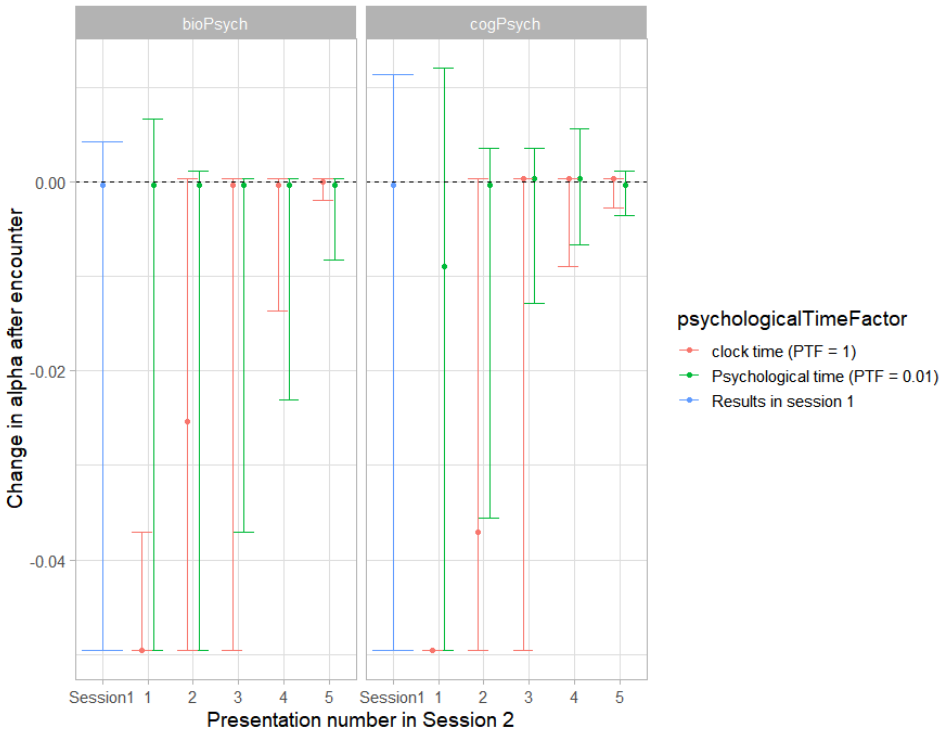

Figure 4. The change in α values after an encounter is plotted over different presentations and sessions. Positive values indicate that the α value was increased after the encounter. The left graph shows the result for the bioPsych dataset and the right sigh for the cogPsych dataset. In blue, the changes in the first session are shown. In red (clock time) and green (PTF = 0.01) the change in α, aggregated by the repetition of the fact in the second session, is shown. The point shows the median value and the error bars show the variance from 0.25-quantile to 0.75 quantile.

presentations in the second sessions were considered because Figure 3 showed that psychological time had the strongest and most direct effect on these encounters. Figure 5 shows histograms of the intervals between the first and second session. The number of sessions over the intervals is right skewed and the number of long intervals is higher for the cogPsych dataset. To make the following analysis more comparable between both datasets only sessions with intervals between sessions of less than 200 hours were considered. This removed 71 (6.6%) sessions in the cogPsych dataset and 44 (5.8%) sessions in the bioPsych dataset.

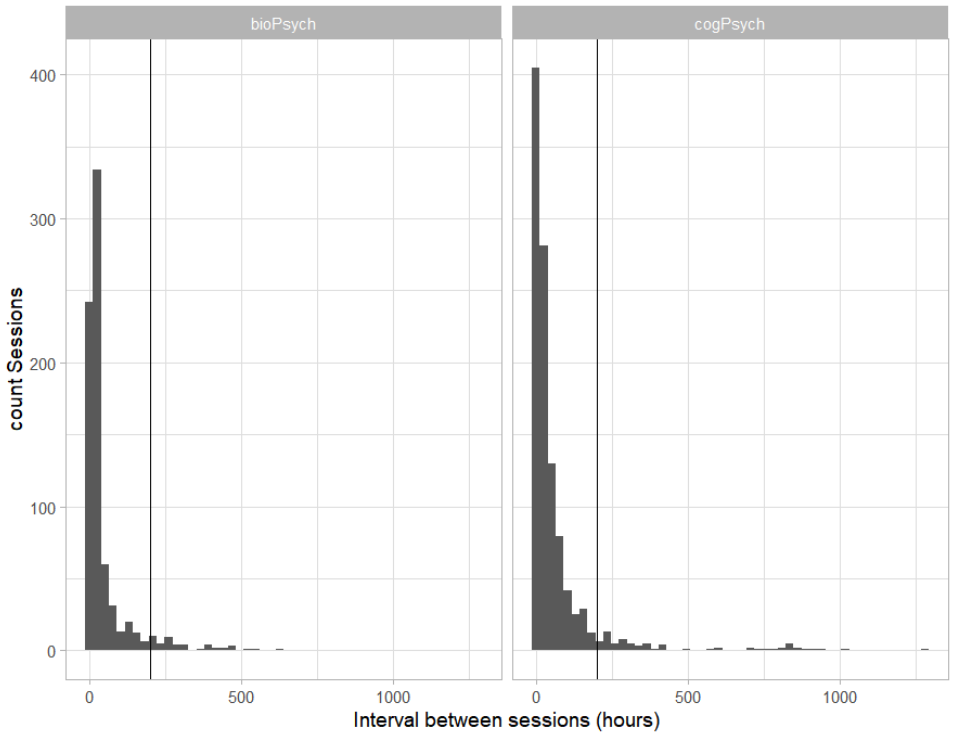

Figure 5. These histograms show the number of learning sessions grouped by intervals between the first and second session for the bioPsych (left) and the cogPsych (right) datasets. The count session indicates the number of sessions of this interval in hours. Blocks are created in ranges of 50 hours. The vertical line indicates 200 hours.

The data was split into five quintiles based on the interval between the first and second session to provide insight into the effect of psychological time over different interval durations. Additionally, in each of these groups the responses were aggregated by the observed responses into five quintiles. Figure 6 shows the results for this aggregation for the cogPsych dataset (the results for the bioPsych dataset are quite similar and are therefore only shown in the Appendix A1). The first facet column shows the distribution of the observed responses. The middle (clock time) and rightmost facet (psychological time) show the predicted response times. For a good fit, the predicted response times should overlap with the observed response times of their respective quintile (shown in different colors) at each interval (shown as rows).

For both cases, clock and psychological time, the quintiles of predicted response times fit the pattern of observed response times: as observed response times increased

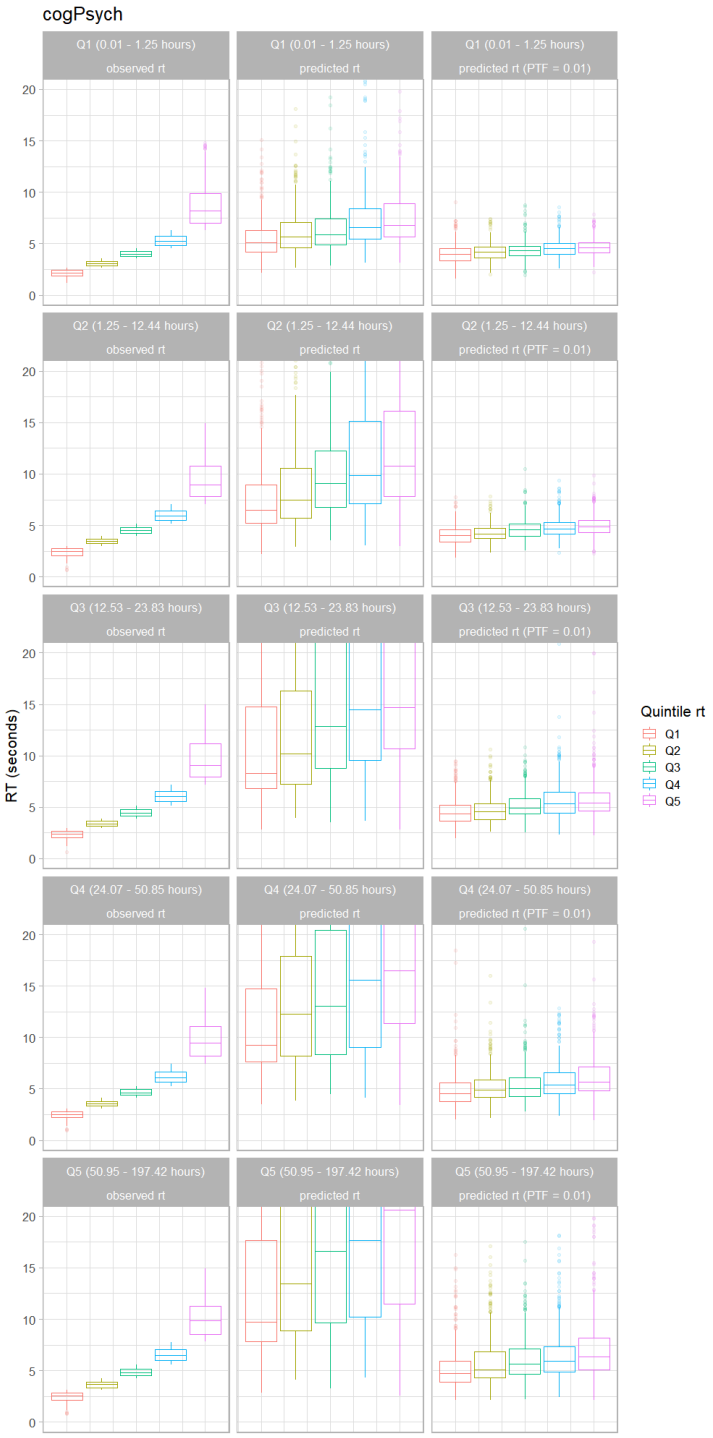

Figure 6. Quintile plot for response times in the cogPsych dataset grouped by intervals (rows) and by observed response times (columns). Boxplots of the observed response times (left), predicted response times in the clock condition (middle) and predicted response times in the psychological time condition (right) (PTF = 0.01) are shown. The whiskers of the boxplots are drawn as in prior plots.

so did the predicted response times. Still both conditions did not fit the observed response times for all quintiles perfectly. Response times were influenced by many different factors, which could not all be considered for the predicted response times (e.g., attention, stress, etc.). Therefore, it should be more important to fit the middle part of the response time distribution rather than the lowest (Q1) or highest (Q5) response times.

The predicted response times between clock time and psychological time differed more with longer intervals between sessions. Comparing these predictions to the observed response times we see that for the psychological time condition the predicted response times were kept within a reasonable range of the observed response times for all intervals. For the clock time condition the responses were increasingly overpredicted with longer intervals between sessions. This indicates that with longer intervals it is more important to include psychological time.

Furthermore, the variance of the predicted response times was increasing for longer intervals between sessions in both conditions. The increase was less in the psychological time condition compared to the clock time condition. The variance of predicted response times was influenced by different decay rates of items. The longer an item could decay between sessions, the higher will be the difference between different decay rates. This was one reason why the variance of predicted response times was higher in both conditions for longer intervals between sessions. The inclusion of psychological time reduced the time between sessions and therefore lead to less variance in the predicted response times. Consequently, the inclusion of psychological time made the predicted response times more reliable especially for longer intervals between sessions.

Up to this point only the results for a fixed PTF of 0.01 were investigated. While Pavlik and Anderson (2005) reported that the PTF tends to be stable over experiments, the following section will analyze if this also holds for the RUGged learning sysetm. Pavlik and Anderson fitted a PTF on intervals that had at least a duration of one day. The datasets used in this thesis also have intervals of shorter time spans (down to a few minutes). Additionally, Pavlik and Anderson (2005) fitted not only the PTF, but also

other parameters (e.g., α, decay) to their ACT-R model, to model the activation of memory traces over time. The RUGged learning system has already predefined functions to fit its parameters during the learning session, making it possible to only focus on fitting the PTF. This allowed us to investigate which PTFs lead to the best fits between observed and predicted response times over a wide range of intervals between sessions. The next section discusses the fitting of psychological time to individual sessions and if the fitted PTFs were affected by the interval between sessions.

Fitting psychological time to whole sessions

This section will first describe a method of estimating the best PTF for each session. Afterwards, we will look at how this best-fitting PTF varied with respect to the interval between sessions. The previous section indicated that a preselected psychological time improved the fit between predicted and observed response time more for longer than for shorter intervals between sessions. The following analysis should provide insight if it is reasonable to assume a constant PTF over different intervals between sessions as it was the case in prior studies (e.g., Pavlik & Anderson, 2005) or if it is necessary to vary the PTF.

Methods

To fit the PTF to a session, a criterion for quantifying the fit of predicted and observed response time was needed. In this thesis, the root-mean-square error (RMSE) was used. The RMSE allowed us to summarize the differences between predicted and observed response times for individual sessions. The RMSE was chosen because it provides an easy to understand and simple measure of fit and is commonly used in the literature (Botchkarev, 2018). Higher RMSE indicates that the predicted response times in the session differed more from the observed response times. Equation 5 shows the calculation for the RMSE.

$$\text{RMSE} = \sqrt{\frac{1}{n} \cdot \sum_{j=1}^{n} (rt_{observed} - rt_{predicted})_j^2} \tag{5}$$

The RMSE was used to investigate which PTF would lead to the best predictions of response times in the second session. The RUGged learning system was used to calculate predicted response times for all encounters in the cogPsych and bioPsych datasets using different PTFs. The chosen PTFs ranged from 0 (no time between sessions) to 1 (clock time). No values below 0 were chosen because these would suggest that the time went back in time in the interval between sessions which seemed unreasonable. On the other hand no values above 1 were chosen because psychological time should shorten the interval between sessions to allow for less decay of activation between sessions. PTFs were chosen in steps of 0.01 from 0 to 1. This provided 101 possible PTFs for each session. After running the RUGged learning system with these different PTFs the data was filtered before the RMSE was calculated. For each session the PTF with the lowest RMSE was selected as being the best PTF. Finally, a generalized additive model (GAM) was used to model these best PTF over different intervals between sessions.

The following steps describe the filtering process. Again only correct responses of learners with responses between 250 milliseconds and 15 seconds were used. Furthermore, only the first presentation of an item in the second session was used. This criterion allows a more direct interpretation of the results. Psychological time influenced the update of the α parameter, though this only impacted the predicted response time after an item has been seen once in the second session. This made it less clear if the improved fit between the observed and predicted response time after the first presentation was based upon the PTF or the update of the α parameter. Additionally, facts that were not seen in the first session were removed because psychological time does not have an effect on them. Lastly, only sessions with an interval between sessions up to 200 hours between the first and second sessions were included. For very long intervals between sessions the distribution of sessions becomes quite sparse. Using only intervals between sessions below 200 hours kept the focus on an interval range with a reasonable density of sessions in both datasets. Applying this criterion reduced the number of sessions from 1076 to 1005 in the cogPsych dataset and

from 765 to 721 in the bioPsych dataset. These sessions were produced from 188 learners in the cogPsych and 298 learners in the bioPsych dataset. After filtering the data the RMSE was calculated for each second session for each student for all selected PTFs. A visualization of the resulting grid can be seen in Appendix B. The average number of trials per session in the cogPsych dataset was 10.07 (standard deviation = 6.13) and in the bioPsych dataset 15.3 (standard deviation = 6.87). In each session the PTF with the lowest RMSE value was chosen to be the best fitting PTF (PTF_{Best}). In case there were multiple lowest RMSE, the mean of these PTFs would have been used, though in both datasets there was always one lowest RMSE.

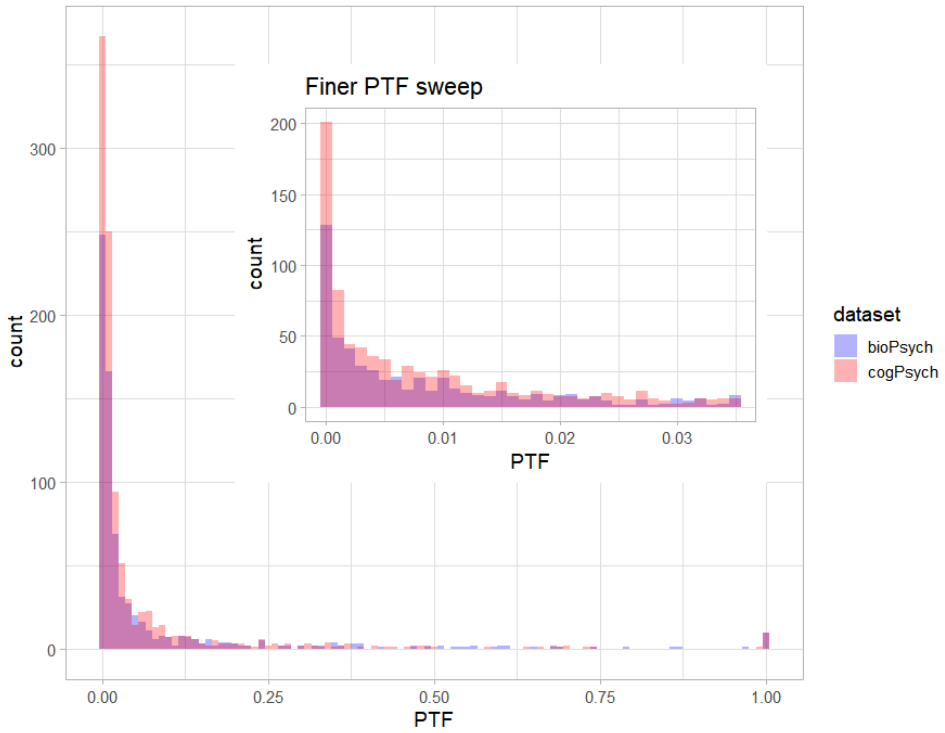

Figure 7. The Figure shows the histograms for the PTF_{Best} for the cogPsych dataset in red and the bioPsych dataset in blue. The embedded plot shows the results if PTFs were selected in steps of 0.001 in the range from 0 to 0.035.

Figure 7 shows the histograms of PTF_{Best} for the cogPsych and bioPsych dataset. In both datasets most sessions had a PTF_{Best} of 0 with a decreasing number of sessions having higher PTF_{Best}. These results fit roughly in the range of prior used PTFs, which ranged from 0.00046 to 0.031 (Pavlik & Anderson, 2003, 2005, 2008; Pavlik et al., 2008).

To get a better picture of this graph an additionally sweep of PTF values in steps of 0.001 up to PTF of 0.035 was taken and the resulting histogram was embedded in Figure 7.

We could now investigate if PTF_{Best} for each session differed over different intervals between sessions between the first and second session. A GAM was used to describe the development of PTFs over intervals between sessions (Lin & Zhang, 1999; van Rij, Vaci, Wurm, & Feldman, 2020; Wood, 2006, 2011). A GAM allows the analysis of non-linear relationships. This was important because there has not been a study on how the relationship between the interval between sessions and the PTF_{Best} should look like. A GAM works well on large datasets for describing data, but has problems with predicting data outside the fitted range. For more information about GAM analysis see Lin and Zhang (1999); van Rij et al. (2020); Wood (2006, 2011).

Results

A GAM was set up for each dataset with the PTF_{Best} as a dependent factor and the interval between sessions as the independent factor (Appendix C). Figure 8 shows the results of fitting the PTF_{Best}. For both datasets the PTF_{Best} was first high and then decreased until roughly half a day later (12 hours). Afterwards, the PTF remained relatively stable. The standard error was wider for longer intervals between sessions because there were fewer sessions with long intervals between sessions.

For both datasets the predicted development of best fitting PTFs seemed to follow a similar trend. Still the predicted PTF tended to be a bit higher for the bioPsych dataset compared to the cogPsych dataset, indicating that there might still be some difference between both datasets. This is not too surprising because both datasets come from different sources with respect to the studying material, the incentive to study and a different distribution of sessions over the interval between sessions. After roughly half a day (12 hours) the PTF_{Best} did not seem to decrease much further in both datasets. One explanation could be that after an interval between sessions of 12 hours students were more likely to sleep between the sessions. Sleep has been shown to have an effect

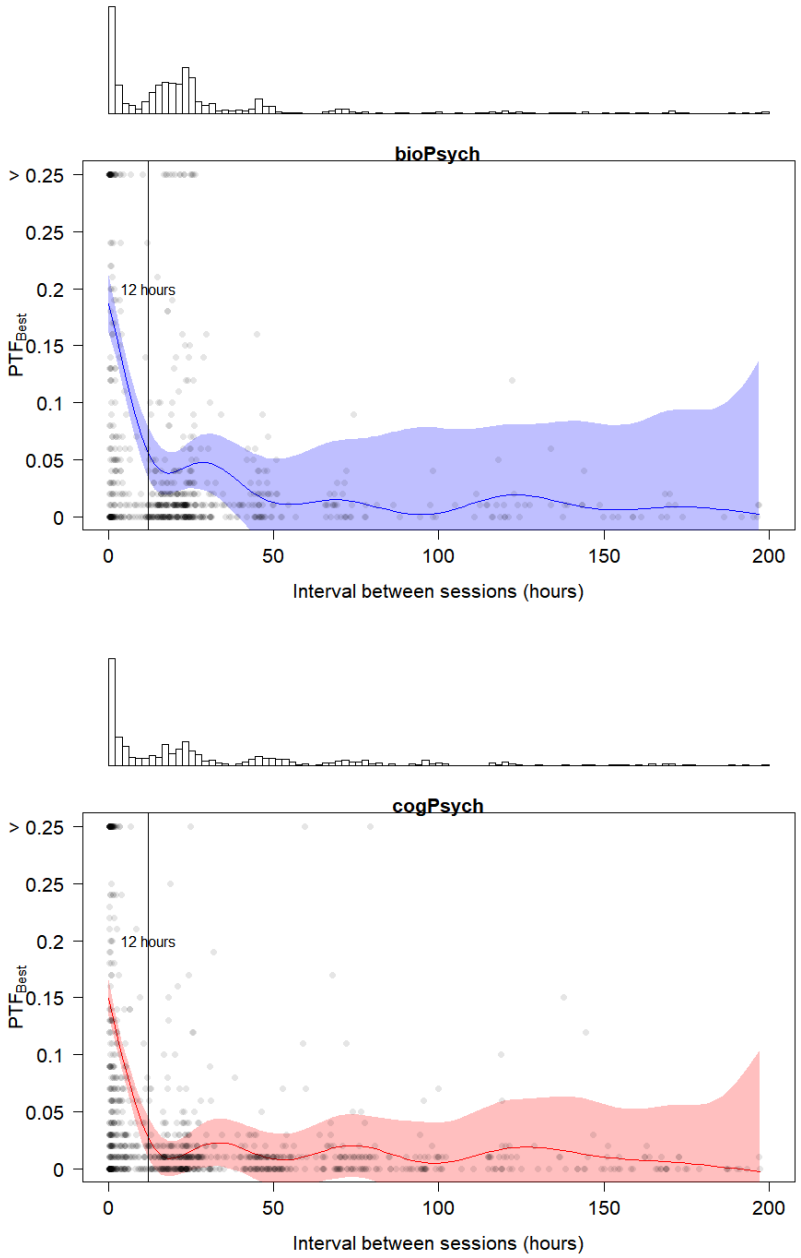

Figure 8. These figures show the fitted GAM on the PTF_{Best} over the interval between sessions between the first to second session. The x-axis shows the interval between sessions in hours. The y-axis shows the PTFs. The upper Figure shows the results for the cogPsych and the lower Figure for the bioPsych dataset. The black points indicate the individual PTF_{Best} for the sessions. Values above 0.25 are shown at >0.25 The transparent ribbons around the graphs indicate the standard errors. The vertical black line indicates the interval between sessions of 12 hours.

on memory consolidation (e.g., Rasch & Born, 2013). The next section will focus on investigating the effect of having a night between sessions on the PTF_{Best}.

Effects of sleep on the optimal PTF

Many studies provided evidence that sleep after learning helps to strengthen newly acquired memories (Diekelmann & Born, 2010; Jenkins & Dallenbach, 1924; Rasch & Born, 2013). This process is called memory consolidation (Petzka, Charest, Balanos, & Staresina, 2020). Initially the idea was that sleep passively improves the retention of memories by protecting against memory interfering events, the current theory is that sleep helps actively with the consolidation of memories by spontaneous reactivations of prior learned material (Rasch & Born, 2013). If sleep helps with memory consolidation, then we would expect that facts decay slower over time after a period of sleep compared to a period of wakefulness.

Still it is not quite clear why memories are better retained after a period of sleep compared to a period of wakefulness. Wamsley (2019) provided evidence that unoccupied waking rest might already have a comparable effect to sleep for memory consolidation. Cordi and Rasch (2020) summarized multiple studies which indicated that effects of sleep might be less robust than general assumed. For example some studies could not replicate that sleep helps to stabilize memories against future interference. In general, it seems that sleep has an effect on memory consolidation, but one should be cautious when assuming a beneficial role of sleep.

Methods

To examine if sleep had an influence on the PTF_{Best} variable, it was necessary to know if learners have slept between the first and second session. Both datasets did not provide such information. The presentation times in the datasets were based upon the location of a central server and did not take the local time of students into account. As most students will probably live around Groningen, we assumed, that all students studied in the time zone of Amsterdam (UTC + 1). This estimate was probably less reliable for students in the bioPsych dataset because they were partly studying during

the lockdown caused by the coronavirus. Based on these time estimates the assumption was made that a student has slept between sessions if the interval between the first and second session included 3 o'clock at night and was at least 7 hours long. The second condition was added because the variable should indicate if learners have slept between conditions: for example if there was only one hour in between sessions, then it would be doubtful that the students have actually slept between the sessions. The 7 hour rule changed three sessions in the bioPsych and one in the cogPsych dataset. The variable *night* is intended to indicate if a student was expected to have slept between the first and second session (*night* = withNight) or not (*night* = noNight). Figure 9 shows the distribution of sessions over the interval between sessions for both noNight and withNight condition. It is important to emphasize that there was not much overlap between both conditions. The factor *night* was included in the GAM as a main effect and as an interaction with the interval between sessions (Appendix C).

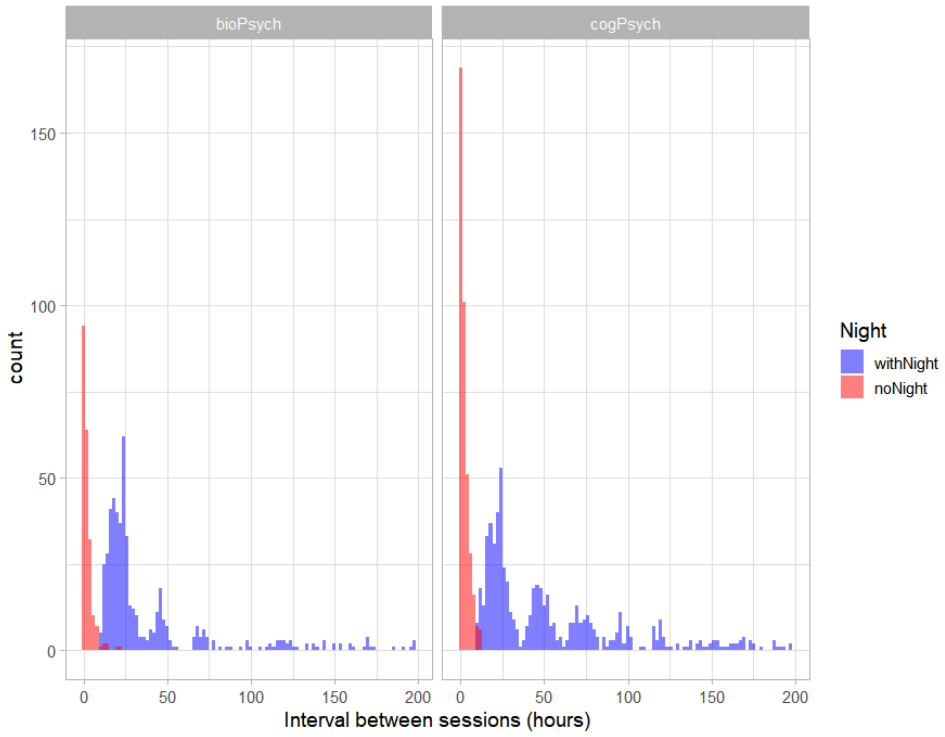

Figure 9. These figures show the histograms for the number of sessions over the interval between sessions in hours for the second session. The bars are two hours long. On the left side the results for the bioPsych dataset and on the right side the results for the cogPsych dataset are shown. The colors indicate the noNight (red) and withNight (blue) condition.

Results

The results of the GAM taking the interval between sessions and *night* into consideration can be seen in Figure 10. A Chi-Square test on the fREML scales of the GAMs with and without *night* indicated that the model that took *night* into consideration fitted for both datasets significantly better than not taking it into consideration (cogPsych: X2(5.00)=16.878, p<.001; bioPsych: X2(5.00) = 14.676, p<.001). This indicates that PTF_{Best} can be better modeled by considering *night*, suggesting that sleep between sessions has an effect on the development of PTF_{Best} over the interval between sessions between first and second session. One problem for this analysis was that the noNight and withNight sessions did not overlap much with respect to the interval between sessions.

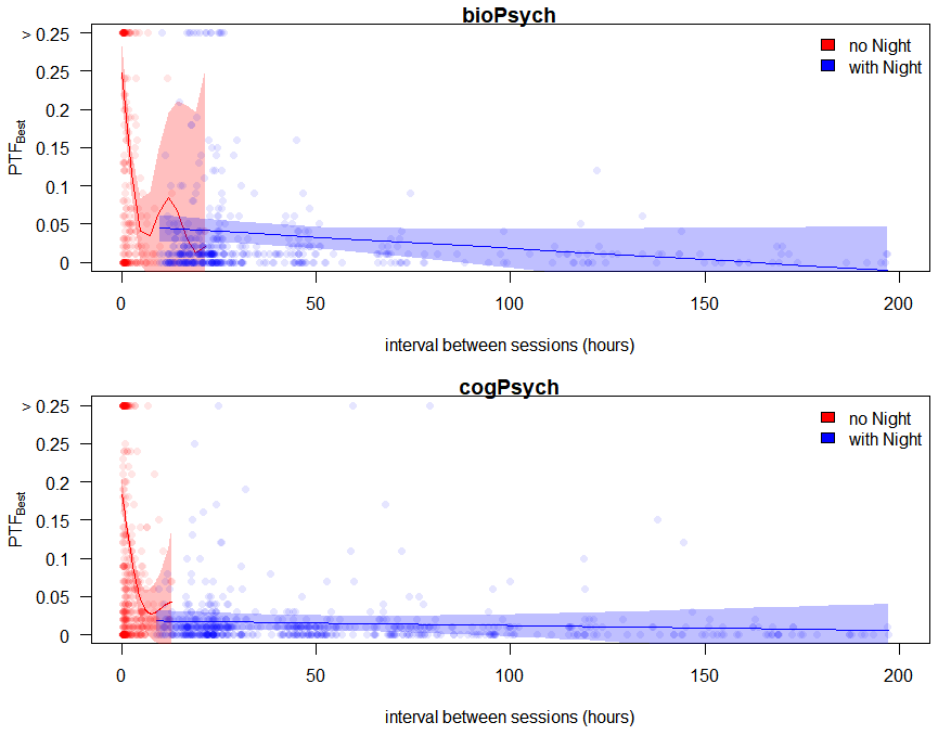

Figure 10. The figures show for the PTF_{Best} over the interval between sessions between the first and second session. The upper graph shows the results for the bioPsych and the lower for the cogPsych dataset. The x-axis shows the interval between sessions in hours. The y-axis shows the PTFs. The red graph shows GAM for noNight and the blue for withNight condition. The transparent ribbons around the graph indicate the standard errors. The PTF_{Best} of all sessions are shown as points in red (noNight) or blue (withNight).

In both graphs in Figure 10 the predicted PTF decreased quickly for the noNight condition. There were few sessions at longer intervals (roughly $>$ 12 hours) for the noNight condition, which created the large standard errors you can see in Figure 10. The condition withNight had a relatively stable PTF_{Best} in both datasets. This might be the result of either having slept or by just having longer intervals between sessions. This is hard to say with the given datasets because there is not much overlap of sessions withNight or noNight with respect to the between session interval. A difference between both datasets was that the PTF_{Best} seemed to decay quicker for noNight within the bioPsych dataset compared to the cogPsych dataset. This might be caused by different

incentives to study between the datasets. In the bioPsych dataset, students needed to reach "Mastery credits" of which they could reach a maximum per day. If participants restudied a session within the same day, they probably had not reached their maximum amount of credits yet, making it more likely that they struggled more with the learning material of that session compared to a session which they only repeated after a day. They might have been more incentives to study right before using the RUGged learning system, to still achieve the mastery points. This would increase the activation of items right before the session. The activation would be higher than expected by the system, which can be accounted for by a lower PTF_{Best}. This could be one reason, why the PTF_{Best} decreased quicker in the bioPsych dataset.

Figure 10 indicates that the PTF_{Best} was higher for the noNight condition compared to the withNight condition. If we can make the assumption that students have slept for the withNight condition and not for the noNight condition then this fits well in the general idea that sleep helps with memory consolidation. If the items are better consolidated due to sleep, then the activation of items should have decayed less. This can be modeled by having a lower PTF.

In summary, there seemed to be a difference in decay of memory traces over intervals between sessions based on if a student had a night between sessions or not. Furthermore, using a constant PTF (as was done in other studies) might not work for all intervals between sessions. It seems to be a reasonable solution for intervals in which a student was expected to have slept (withNight). On the other hand, if the intervals between sessions were shorter and the student could not have a night between sessions, then it might be more reasonable to use a varying PTF. The following section will present and test different functions for modeling the PTF_{Best} over the interval between sessions.

A simpler model for the development of the PTF

Is there an underlying function for how the PTF_{Best} develops over the interval between sessions? Prior papers used one PTF for different intervals though they only

looked at intervals of one day or more (e.g., Pavlik & Anderson, 2005). The prior section indicated that this might be reasonable, if learners were expected to have a night between sessions. On the other hand, the PTF_{Best} seemed to decrease if learners had no night between learning sessions. In this section, different functions will be examined to see if they can be used to describe and predict the psychological time over the interval between sessions. A simpler function would be helpful for integrating a PTF into the RUGged learning system because the GAM is quite complex, which can cause problems for predicting PTFs outside the range of intervals between sessions the GAM was fitted on. Furthermore, there is likely more general scientific interest in modeling memory decay over time periods exceeding the intervals used in this thesis.

Methods

Functions for development of the PTF. Functions that were tested for describing and predicting the PTF_{Best} over intervals between sessions were a constant function, a linear function, a power function and an exponential function (Table 2). The constant function was the default option because it is the simplest function and prior papers (e.g., Pavlik & Anderson, 2005) used one PTF over different intervals between sessions. The GAM in Figure 8 showed a decreasing PTF_{Best} over intervals between sessions for which a more complex function might fit better. For this reason the power and exponential function were included. The linear function was included as a trade-off between complex and simple functions. We only fitted these functions because they seemed to be reasonable candidates based on the results from the GAM functions.

The prior mentioned functions were fitted using the nonlinear least-squares (nls) method (Bates & Watts, 1988) with the Levenberg-Marquardt modification (Moré, 1978) using the minpack.lm package (Elzhov, Mullen, Spiess, & Bolker, 2016). This method uses an extension of the Gauss-Newton algorithm to find the best fitting parameters of the selected functions for a given dataset. The Levenberg-Marquardt modification was used to improve convergence. The nls requires the selection of starting values for the individual parameters (a, b, c), which should be a reasonable estimate.

Table 2 shows the used starting parameters. The nls was allowed to iterate over the data up to 200 times to converge on parameters.

Table 2

Four different tested functions with their equations and their chosen starting values for the nls parameter estimation. The variable interval represents the interval between sessions.

function type	Equation	a	b	c
constant	$PTF = a$	0.01	NA	NA
linear	$PTF = a + b \cdot \text{interval}$	0.01	0.25	NA
power	$PTF = a + b \cdot \text{interval}^c$	0.01	0.25	0.2
exponential	$PTF = a + b \cdot c^{\text{interval}}$	0.01	0.25	0.2

Fit within dataset. Bootstrapping was used to test which function can robustly fit the PTF_{best} over different intervals between sessions. Functions were fitted individually to the noNight (214 bioPsych and 378 cogPsych PTF_{Best} values) or withNight condition (507 bioPsych and 627 cogPsych PTF_{Best} values) for both the bioPsych and cogPsych dataset separately. This difference was chosen because the GAM in the prior section indicated that there was a difference for these two conditions. The datasets were resampled 1000 times with replacement. The four different functions from Table 2 were fitted on these resampled datasets using nls. The exponential function in the withNight condition was not able to converge on parameters for all bootstrapped datasets (non-convergence for 83 (cogPsych) and 66 (bioPsych) of 1000 resampled datasets). For all other cases convergence was reached.

To assess how well the functions fitted the resampled data, the Akaike information criterion (AIC) was calculated using the standard AIC equation [1] of the R (R Core Team, 2019). The AIC is an estimator for the out-of-sample prediction error. The AIC provides an indication for model fit while penalizing model complexity. It estimates model fit relative to a given dataset, so only AIC scores on the same dataset can be

[1] The equation can be found in Appendix D.

compared against each other. Lower AIC scores indicate a better function type for this dataset.

To still allow comparison of fits across the bootstrapped samples, the AIC values were rescaled to Δ_i following Burnham and Anderson (2004). Δ_i showed for a given dataset how many AIC values worse the function was compared to the best fitting function for that dataset. Δ_i was calculated according to Equation 6 for each resampled dataset.

$$\Delta_i = AIC_i - AIC_{min} \tag{6}$$

Within each dataset the lowest AIC value of the four functions was subtracted from all AIC values. Higher Δ_i for a function indicated that it was less likely that this function was the best fitting function. As a rule of thumb, $\Delta_i > 10$ indicates essentially no support for being the best model, $4 < \Delta_i < 7$ indicates considerable support and $\Delta_i < 2$ indicates substantial support (Burnham & Anderson, 2004).

Fit between datasets. Lastly, we wanted to evaluate how well these functions generalize to the other dataset: how well can you predict PTF_{Best} values from the bioPsych dataset using functions fitted to PTF_{Best} values from the cogPsych dataset and vice versa? For this the same bootstrapped functions as before were used but now their fit was evaluated based on how well they can predict the PTF_{Best} from the other dataset. For example the exponential function was fitted to 1000 bootstrapped datasets of the noNight condition of the bioPsych dataset. These 1000 bootstrapped functions were then tested on the noNight condition of the cogPsych dataset. Results were quantified by calculating the root-mean-square error (RMSE) for every function based on the predicted PTF values and the observed PTF_{Best} values for all sessions. The RMSE was used instead of Δ_i because functions should not be penalized for model complexity anymore if they are tested on a new dataset.

Results

Bootstrapped predictions. The fitted functions based on the bootstrapped datasets for the noNight condition can be seen in Figure 11. In general the predictions seem to be relatively stable for all sessions, meaning that for each function different resamplings of the dataset produce similar patterns. The conclusions will therefore probably not depend on individual influential data points. One exception was the power function for the bioPsych dataset noNight condition. In this case, there seemed to be two sets of good fits in the data, which might suggest that not one power function could fit different resamples of the data well.

For the withNight condition (Figure 12) the predicted PTFs were relatively similar for all functions. There were two patterns in the fitted exponential functions: either increasing or decreasing PTF_{Best} over the interval between sessions with the biggest difference within shorter intervals. This indicates that for the exponential function there was not a robust pattern of decreasing or increasing PTF_{Best} over intervals between sessions.

Some functions predicted negative PTF. Negative values are problematic because there was no session in which the PTF_{Best} was negative in the datasets. Furthermore, it seemed unreasonable to assume negative values because this would mean that the activation of facts increased the longer you wait between sessions. The strongest negative values were for the exponential function, which increased over intervals in the withNight condition in the first few intervals and for the linear function for later intervals. Especially the negative values at longer intervals might be a problem for other studies that use a PTF for intervals longer than 200 hours. Constraining the range in which the functions can predict the PTF to for example 0 to 1 could alleviate this problem. This was not done in this study because most predicted PTFs were already in a range between 0 and 1.

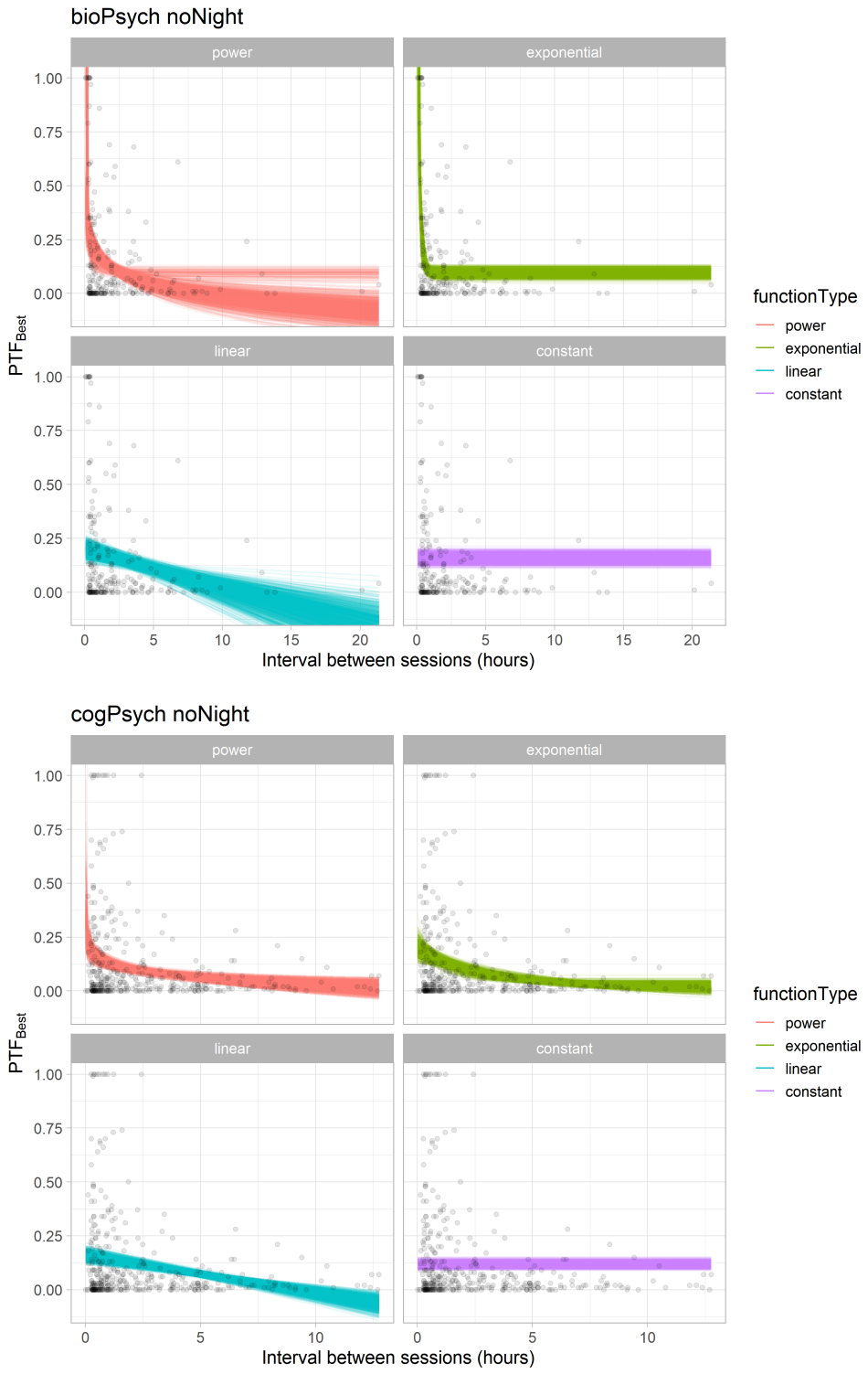

Figure 11. 1000 bootstrapped fits of the four equations for predicted PTFs over intervals between sessions in hours for the noNight condition. The top figure shows the results for the bioPsych and the bottom one for the cogPsych dataset. The facets show the predicted PTFs for the different function types. The PTF_{Best} are shown as points.

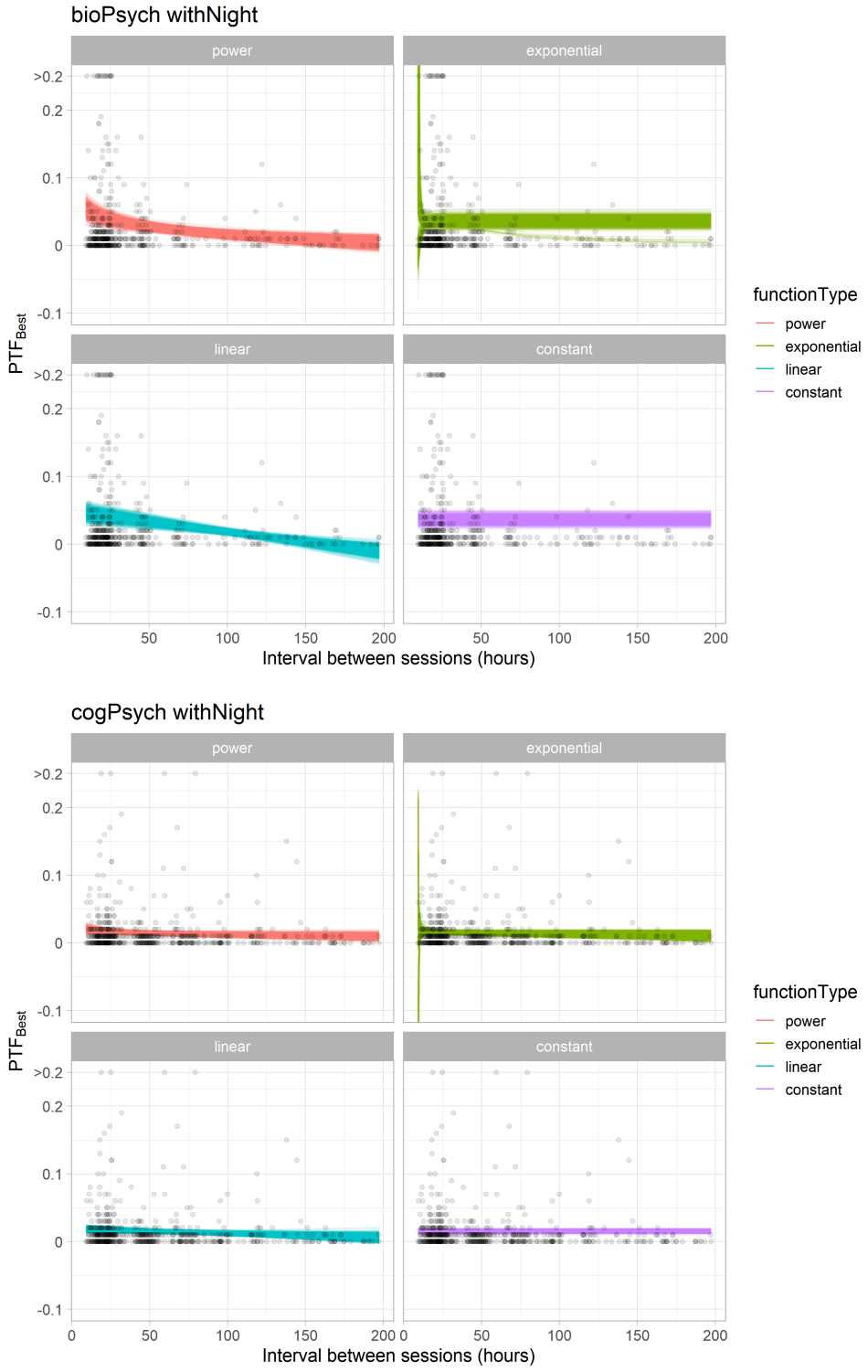

Figure 12. 1000 bootstrapped fits of the four equations for predicted PTFs over intervals between sessions in hours for the withNight condition. The top figure shows the results for the bioPsych and the bottom one for the cogPsych dataset. The facets show the predicted PTFs for the different function types. The PTF_{Best} are shown as points. Values above 0.2 are shown at y-axis as >0.2.

Fit within dataset. Figure 13 shows how well the functions fitted the bootstrapped datasets in the noNight condition. In both datasets the exponential function tended to have the lowest Δ_i values and the constant function the highest Δ_i values. The difference seemed to be robust over multiple resamples of the datasets. Assessing how often each function type was the best function for a bootstrapped dataset ($\Delta_i = 0$) we see that the exponential function was the best in the noNight condition in 95.7% (bioPsych) and 86.8% (cogPsych) of cases. Taking the rule of thumb into consideration (substantial support: Δ_i<2, no support Δ_i>10) it indicates that there was essentially no support for the constant function. Furthermore, we see that the exponential function in both datasets tended to have Δ_i values below 2 indicating that there was considerable support for this function. For the linear and power function the results were less straightforward. In the bioPsych dataset, their Δ_i values tended to be above 10 and in the cogPsych dataset the values tended to be below 10. This indicates that in the noNight condition they were clearly not the best function types for the cogPsych dataset, while they were reasonable function types in the bioPsych dataset.

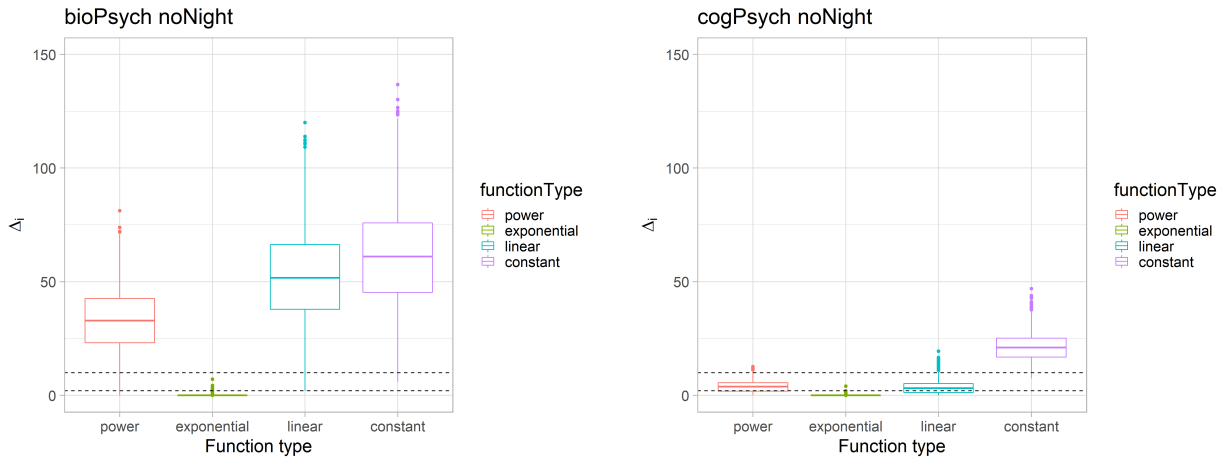

Figure 13. Δ_i is shown over different function types for the bioPsych (left) and cogPsych (right) dataset for the noNight condition. The dashed lines indicate Δ_i of 2 and 10 (rule of thumb). The whiskers of the boxplots are drawn based upon the largest observed point, which falls within the range of $1.5 \cdot IQR$ above the 75% quantile and the lowest observed point, which falls within the range of $1.5 \cdot IQR$ below the 25% quantile.

Figure 14 shows how well the functions fitted the bootstrapped datasets in the withNight condition. For the withNight dataset the nls was not able to find parameters for the exponential function for all datasets (not converged in cogPsych 79 and bioPsych 61 datasets of 1000). This indicates that the data was more difficult to fit with an exponential function, highlighting that the exponential function was probably not a good function for the withNight condition. For the bootstrapped datasets, which could be fitted, we see that it tended to have a reasonable fit ($\delta_i < 10$), though it did not have strong support for being the best function ($\delta_i > 2$). The linear function tended to have the lowest Δ_i ($\delta_i < 2$) and had the most Δ_i of 0 (bioPsych: 80.5%, cogPsych: 66.9% of bootstrapped datasets). Still Δ_i tended to be below 10 for all functions indicating that for most resampled datasets all function types might be reasonable suggestions. While the linear function tended to have the lowest Δ_i values, the constant function was not clearly worse. As prior studies only used one PTF over different intervals between sessions and we saw that the liner function can suggest unreasonable (negative) PTFs, it might be reasonable to say that the constant function seems to be sufficient to account for the development of PTF in the withNight condition.

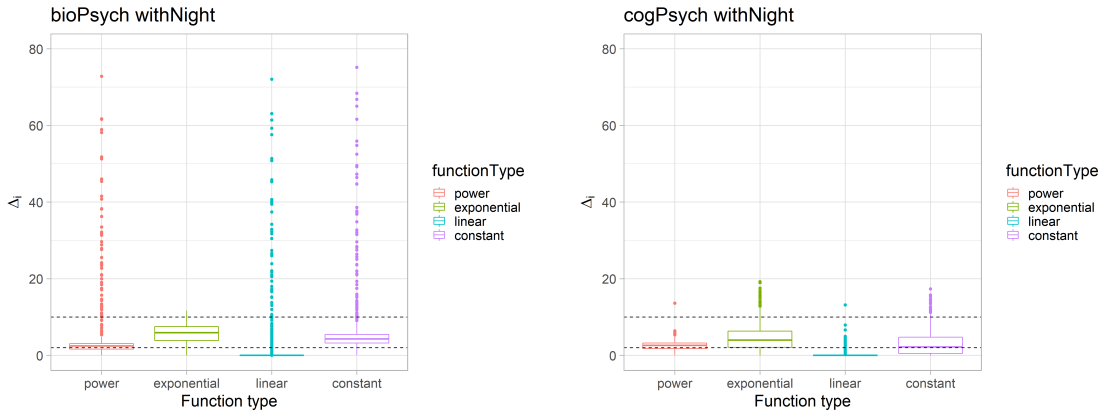

Figure 14. Δ_i is shown over different function types for the bioPsych (left) and cogPsych (right) dataset for the withNight condition. The dashed lines indicate Δ_i of 2 and 10 (rule of thumb). The whiskers of the boxplots are drawn based upon the largest observed point, which falls within the range of $1.5 \cdot IQR$ above the 75% quantile and the lowest observed point, which falls within the range of $1.5 \cdot IQR$ below the 25% quantile.

Fit between datasets. Figures 15 and 16 show RMSE values for fitting the bootstrapped functions from the cogPsych dataset on the bioPsych dataset and vice versa for the noNight and withNight condition, respectively. Based on the results from fitting functions within the same dataset the assumption was that the exponential function would fit the best for the noNight condition and that that there was not much difference between the fit of different function times for the withNight condition. We wanted to see if these results generalize to another dataset.

Figure 15 shows the results for the noNight condition. In the bioPsych dataset the power function had the lowest RMSE, indicating that it predicted the PTF values best. In the cogPsych dataset the linear function had the lowest RMSE. In this dataset the exponential function also seemed to have clearly the worst fit (highest RMSE), indicating that it did not generalize well to the cogPsych dataset. Still there was not clearly a function type that predicts the PTF best for both datasets.

Figure 16 shows the result for the withNight condition. Here the power and linear function types had the lowest RMSE values in both datasets, indicating that they tended to generalize best to the new dataset, though the differences to the other function types were quite small, indicating that there might not be one function that fitted clearly better than the other functions. In general the results do not allow to clearly state that one function type generalized better than the others to a new dataset.

Within the noNight condition the fitted exponential function did not generalize best to the new dataset. For the cogPsych dataset it even fitted the worst. This might indicate that it was overfitted or that the functions were dataset specific. For the noNight condition, the RMSE did not differ much from each other, indicating that there was not one function type that clearly generalized best to a new dataset. In a practical sense this means, if you want to use a function to model the development of PTF over intervals you might be able to describe the development of PTF over the interval between sessions by using an exponential function (noNight) and constant function (withNight), but these results will not necessarily generalize well to new datasets. Possible reasons for this will be discussed in the discussion section.

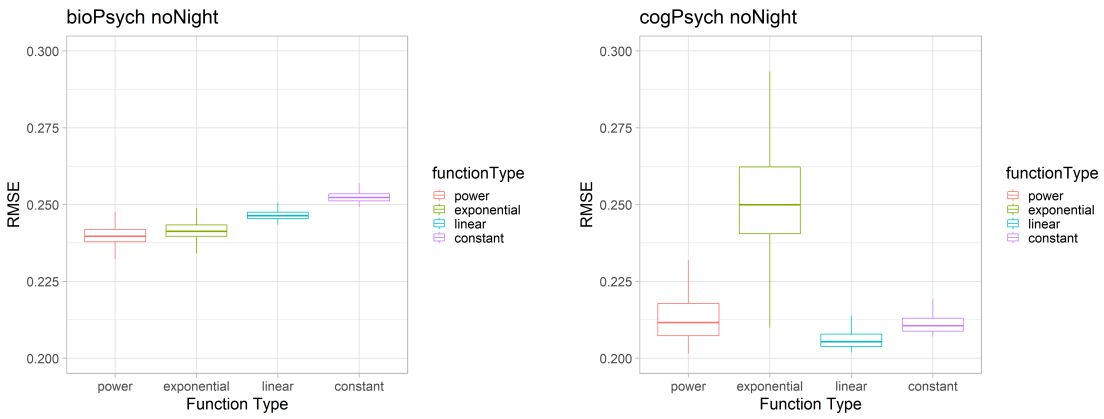

Figure 15. The RMSE of the fit between predicted PTF and PTF_{Best} is shown over different function types. The left graph shows how well the functions trained on the resampled cogPsych dataset fit the bioPsych dataset. The right graph shows it the other way round. These graphs show the results for the noNight condition. RMSE values outside the whiskers were removed to improve the readability of the graphs.

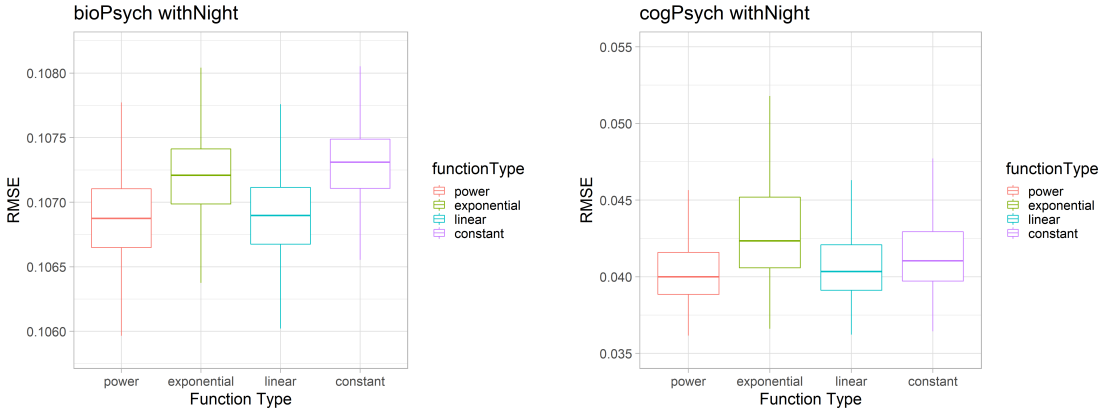

Figure 16. The RMSE of the fit between predicted PTF and PTF_{Best} is shown over different function types. The left graph shows how well the functions trained on the resampled cogPsych dataset fit the bioPsych dataset. The right graph shows it the other way round. These graphs show the results for the withNight condition. RMSE values outside the whiskers were removed to improve the readability of the graphs.

Discussion

In this thesis we investigated the effects of extending an adaptive learning system (RUGged learning system) with the construct of psychological time. Here, the interval

between sessions was scaled by a psychological time factor (PTF), to account for the slower decay of activation over time between learning sessions compared to within learning sessions. At the beginning of the thesis a fixed PTF was used. Afterwards, we fitted individual PTFs to learning sessions and finally examined their development over the interval between sessions. The results of this study support the hypothesis that one should account differently for the time between compared to the time within learning sessions to model the activation of facts. We have shown that the inclusion of psychological time improved the response time predictions of the RUGged learning system to better fit the observed response times in two naturalistic datasets. The strongest improvement was on the first presentation of encounters in the second session. For a fixed PTF the improvement increased with longer intervals between sessions. Furthermore we found that the PTF for a learning session should be informed by the interval before that session. With longer intervals the optimal PTF decreased and stayed relatively constant when there was a night between learning sessions.

These findings are consistent with prior studies that found that the activation decays slower between learning sessions compared to within learning sessions (e.g., McBride & Dosher, 1997) and that psychological time can account for this (e.g., Pavlik & Anderson, 2003, 2005). Contrary to the prior method we have shown that you need to account for the passage of time in a more refined way than just a single PTF. While Anderson et al. (1999) modeled the inclusion of psychological time as a discrete shift at the end of the session, they also mentioned that it would be possible that this shift is more continuous rather than discrete. In other words the time at which the activation of facts decays might slow down more gradually after a learning session. The datasets used in this thesis had a variety of different intervals between sessions and were collected in a naturalistic setting. This allowed us to map out the development of PTF over intervals in more detail compared to what was done in prior studies. We showed a tendency of decreasing PTF values with increasing intervals. This points at that there is a more continuous rather than discrete slowdown of psychological time after a learning session.

The strongest difference between psychological time and clock time was for the

first presentations of facts in the second session. While there was still some difference on the second presentation, there was barely any difference anymore on later presentations. For the analysis of the optimal PTF, we only looked at fits for the first presentation. Potentially, the second presentation should also be taken into consideration, though this was not done. It is unclear where the discrepancy on the second repetition originates from. For example if the predicted response time for the second repetition is too high then this could be because the time between sessions is not correctly accounted for or because the update of α is not working well together with psychological time.

The inclusion of psychological time is also interesting with respect to the α parameter of the RUGged learning system. If no psychological time is included, then the RUGged learning system compensates for the slower decay between sessions by reducing its α parameter. Including psychological time reduced the amount by which the α parameter needs to be changed, this should lead to more stable α values between sessions and allow the system to better fine-tune to the individual decay rates of the learner/ item combination.

As a pragmatic solution for the inclusion of psychological time, it seems already feasible to just use a single low PTF value (for example 0.01). As shown in this thesis this value will probably improve the predicted response times quite well over all intervals. While it seems that the optimal PTF tends to be higher than 0.01 for shorter intervals, psychological time was also not as impactful at shorter intervals compared to longer intervals, making it potentially less important to use the optimal PTF for shorter intervals.

In this study we have shown that psychological time changes the activation function and that this improves the predictions of the response times. Modeling the response times of the learners correctly is an important intermediate goal for a fact learning system, but the end goal is to improve the learning outcomes. If psychological time is applied to the RUGged learning session during learning, then the change for the activation values would also change the presentation order of encounters, which could influence the learning process. A future study should be conducted to see if the

inclusion of psychological time in the RUGged learning system actually helps to improve learning outcomes. A potential setup of such a study could be: students have two learning sessions with the RUGged learning system separated by one day. For one group the time between sessions is scaled by a PTF of 0.01 and for the other the interval is not scaled. After the second session, both groups are tested on the learned material. As this thesis showed the strongest differences between psychological time and no psychological time within the first few repetitions within a session, it would be useful to keep the second learning session in this experiment relatively short. If psychological time improves the learning outcomes (for example number of facts gotten right on the final test), then a further study might want to vary the length of the learning sessions and the time between the learning sessions.

Within a dataset (bioPsych and cogPsych) the development of PTF over the interval between sessions in the noNight condition was best described by an exponential function. Unfortunately, the parameterization of this exponential function did not generalize well to the other dataset. This could be because the exponential function was overfitted. We tried to account for this by penalizing model complexity. Furthermore, the bootstrapped exponential functions showed relatively similar patterns within each dataset, indicating that the fit was probably not influenced by individual influential points. On the other hand it might also be that there actually is a difference in how the PTF develops over the interval between those datasets. The datasets differ in several ways. Some influential factors might include different learning materials and different incentives. Students in the bioPsych dataset needed to get mastery credits, which were awarded on an accuracy-based and effort-based criterion up to a maximum of one mastery credit in 24 hours per chapter. Students in the bioPsych dataset might for example be more likely to study between sessions to get their mastery credits quicker. Additionally, in the bioPsych dataset in the first 24 hour interval, easier chapters might be underrepresented. If a chapter was simple enough that students could already get their maximum one mastery credit with one session, then they will probably only restudy this chapter, if they can gain the next mastery credit (after 24 hours). In

summary, it seems that within a domain (for example cogPsych or bioPsych) the exponential function can be used to describe the development of PTF over intervals within the same day. The precise parameters of such a function can differ between domains. A future study might provide insight into which factors influence the parameters of such a function.

Even if we know which factors influence the PTF for a session, the learning system might not always have access to these information. Furthermore, even within a learning domain a single function could not capture every PTF for all sessions. A different approach for finding the optimal PTF might be to adjust the PTF at the beginning of a learning session to the responses of a learner. How could this look like? For example the system recognizes that the learner has responded quicker to the first facts in the second session. It therefore adjusts the PTF to a lower value. This will then influence the predictions for all following encounters. To get a reasonable starting PTF, the exponential function (for intervals with a night) or a constant function (for intervals with no night) could be useful. Based on the interval between sessions a starting value can be selected. These starting values are then continuously adjusted within a session based on the responses of the learner. It is important to fit the PTF on multiple responses because response times can fluctuate a lot. A potential problem is the interaction between fitting the α and the PTF on the same encounters. If a learner responses quicker than expected, decreasing the PTF or α would both improve the predicted response time for this encounter. Here, it is important to point out that we found the biggest improvement of the PTF for the first presentations of items in a session. A possible solution could be to first adjust the PTF and only after a few encounters adjust the α again. At which point within the session this shift should occur, for example after one minute or after the first five correct items, needs to be investigated in a future study.

On the one hand we have shown that the inclusion of a PTF improves the predictions of the RUGged learning system, on the other hand it is not quite clear what this factor actually models. The first idea in Anderson et al. (1999) was that

psychological time accounts for the number of memory interfering events, but there are likely more other influences affecting this factor. There is some evidence that interfering events are less harmful with longer consolidation time (Verhoeven & Newell, 2018) and facts might get better consolidated over time. After a session there might not only be fewer interfering events, but these events might also become less and less disruptive. Related to this idea of consolidation, we found that PTF is lowest after we expected learners to have slept between sessions, indicating that sleep is also influencing the PTF. As sleep has been shown to help with memory consolidation, one interpretation would be that sleep reduces the PTF by improving memory consolidation. Lastly, students might have studied the learned material between learning sessions without the RUGged learning system, which would also increase the activation of facts at the beginning of a second learning session. At this point PTF does not have a strong theoretical background and should be seen as a placeholder for one or more underlying effects that influence the decay of activation between sessions. Yet psychological time holds some practical value because it clearly improved the predictions of the learning system. For the future PTF should be related to factors in the interval between sessions because this would help to make the concept more theoretical sound.

Instead of PTF, there are possible alternative ways to account for the slower decay of activation between learning sessions. For example it might be that at different time scales different processes affect the learned memories (Verhoeven & Newell, 2018). Newell, Mayer-Kress, Hong, and Liu (2009) modeled the motor learning of their learners by a function that keeps track of transient learning (improvement within a session) and persistent learning (improvement throughout long intervals). This two-time scale model assumes that there are at least two underlying processes which affect the learning for different time scales. This approach might help to get a better understanding of underlying memory processes. The problem is that this approach is more complex because you need to fit a second function to the data. Furthermore, it did not produce better results than the use of psychological time (Newell et al., 2009). Lastly, in this thesis we showed that the PTF is relatively constant after a night but decreases

beforehand. One explanation for this is that memories are better consolidated after the night. Newell et al. (2009) fitted the two-time scale model to datasets, where learning sessions were on different days. It could be that the persistent learning only captures the memory development after learners have slept and therefore the memory of facts is well consolidated. In this case their current approach would fall short for intervals between sessions in which the memories are not well consolidated yet. The focus of this thesis was more on what is practical to improving the predictions of an adaptive learning system rather than having a theoretical sound construct, which explains the underlying processes.

Broader implications

Typically studies focus on the distribution of practice within or between sessions (Verhoeven & Newell, 2018). Psychological time could help to bring both together. For example the RUGged learning system tries to present items right before they are forgotten within a session. Using psychological time one might also make an informed decision on when to repeat the learning session again. For example to repeat the learning session again, just before the learner is expected to have forgotten the prior learned facts.

Furthermore, the inclusion of psychological time might allow a learning system to create the optimal learning environment for the learner more quickly. With psychological time the predicted response time fitted the observed response times already well for the first time items were repeated within the second session. Imagine you have a few minutes time to study. Then you want to have an effective learning session right away and not give the system a few repetitions to adjust to you. With psychological time a learning system would be ready to be used more quickly.

It is also interesting to relate the results to the research on sleep. In this thesis we saw that the optimal PTF for a session was lowest after we expected students to have slept between sessions. Rasch and Born (2013) mention the idea, that the brain is optimized for memory consolidation during sleep while being optimized for memory

acquisition during the time of wakefulness. This fits well with our findings. If memories are getting better consolidated, then they should also decay slower over time and we showed that the PTF is lower after a night of sleep. There are some limitations for these findings. As the datasets were collected under naturalistic settings, there was no control for interval length and there was only a small overlap of intervals in which students have slept or have not slept. Even though the GAM including the variable *sleep* modeled the development of PTF_{Best} better than without this variable, we cannot fully exclude that with longer intervals the learned material is consolidated better and sleep is just correlated with higher intervals. While it seems to make sense that facts are better consolidated after learners have slept and therefore the PTF is lower, an experiment needs to be conducted to see how exactly PTF is related to sleep and to refine this analysis further. Here it would be essential to know when learners have slept. Additionally it needs to be controlled that there is a sufficient amount of sessions with the same interval length between sessions in which learners have slept or not. This can be difficult due to the circadian rhythms.

Conclusion

In summary, the present study contributes to a growing body of evidence that you need to model activation of memories over time differently within a learning session compared to between learning sessions. Using psychological time can account in a relatively simple way for this difference. While future research is needed to validate these findings, this thesis provides clear support for the usefulness of psychological time for modeling the activation of facts over time.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* (Vol. 3). Oxford University Press.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*(5), 1120-1136. doi: 10.1037/0278-7393.25.5.1120

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408. doi: 10.1111/j.1467-9280.1991.tb00174.x

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications* (Vol. 2). Wiley New York.

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *Interdisciplinary Journal of Information, Knowledge, and Management*, *14*, 45-79. doi: 10.28945/4184

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, *33*(2), 261-304. doi: 10.1177/0049124104268644

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. doi: 10.3758/BF03202713

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. doi: 10.1037/0033-2909.132.3.354

Cordi, M. J., & Rasch, B. (2020). How robust are sleep-mediated memory benefits? *Current Opinion in Neurobiology*, *67*, 1-7. doi: 10.1016/j.conb.2020.06.002

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In *Psychology of learning and motivation* (Vol. 53, pp. 63–147). Elsevier.

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, *11*(2), 114-126. doi: 10.1038/nrn2762

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805. doi: 10.1037/0021-9010.84.5.795

Edge, D., Fitchett, S., Whitney, M., & Landay, J. (2012). Memreflex: adaptive flashcards for mobile microlearning. *14th international conference on Human-computer interaction with mobile devices and services*, 431-440.

Elliott, S., & Anderson, J. R. (1995). The effect of memory decay on predictions from changing categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 815-836. doi: 10.1037/0278-7393.21.4.815

Elzhov, T. V., Mullen, K. M., Spiess, A.-N., & Bolker, B. (2016). minpack.lm: R interface to the levenberg-marquardt nonlinear least-squares algorithm found in minpack, plus support for bounds [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=minpack.lm` (R package version 1.2-1)

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: The role of retrievability. *Quarterly Journal of Experimental Psychology*, *65*(5), 962-975. doi: 10.1080/17470218.2011.638079

Jenkins, J. G., & Dallenbach, K. M. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology*, *35*(4), 605-612. doi: 10.2307/1414040

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. doi: 10.1126/science.1152408

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society (B)*, *61*(2), 381-400. doi: 10.1111/1467-9868.00183

McBride, D. M., & Dosher, B. A. (1997). A comparison of forgetting in an implicit and explicit memory task. *Journal of Experimental Psychology: General*, *126*(4), 371-392. doi: 10.1037/0096-3445.126.4.371

Moré, J. (1978). The levenberg-marquardt algorithm: Implementation and theory. In *Numerical analysis* (Vol. 630, pp. 105–116). Springer. doi: 10.1007/BFb0067700

Newell, K. M., Mayer-Kress, G., Hong, S. L., & Liu, Y. T. (2009). Adaptation and learning: Characteristic time scales of performance dynamics. *Human movement science*, *28*(6), 655-687. doi: 10.1016/j.humov.2009.07.001

Nijboer, M. (2011). Optimal fact learning: Applying presentation scheduling to realistic conditions. Groningen, The Netherlands: Unpublished Master's Thesis, University of Groningen.

Pavlik, P., & Anderson, J. R. (2003). An act-r model of the spacing effect. In F. Detje, D. Doerner, & H. Schaub (Eds.), *In proceedings of the fifth international conference on cognitive modeling* (p. 177– 182). Bamberg, Germany: Universitaets-Verlag Bamberg.

Pavlik, P., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation based model of the spacing effect. *Cognitive Science*, *29*(4), 559–586. doi: 10.1207/s15516709cog0000_14

Pavlik, P., & Anderson, J. R. (2008). Practice and forgetting effects on vocabulary memory: An activation based model of the spacing effect. *Journal of Experimental Psychology: Applied*, *14*(2), 101–117. doi: 10.1037/1076-898X.14.2.101

Pavlik, P., Bolster, T., Wu, S. M., Koedinger, K., & Macwhinney, B. (2008). Using optimally selected drill practice to train basic facts. In *International conference on intelligent tutoring systems* (p. 593-602). Berlin, Heidelberg.

Petzka, M., Charest, I., Balanos, G. M., & Staresina, B. (2020). Does sleep-dependent consolidation favour weak memories? *PsyArXiv*. Retrieved from `10.31234/osf.io/q4wnv` doi: 10.31234/osf.io/q4wnv

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological reviews.*.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. doi: 10.1016/j.tics.2010.09.003

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, *8*(1), 305-321. doi: 10.1111/tops.12183

Sense, F., Meijer, R. R., & van Rijn, H. (2018). Exploration of the rate of forgetting as a domain-specific individual differences measure. *Frontiers in Education*, *3*(112). doi: 10.3389/feduc.2018.00112

Sense, F., van der Velde, M., & van Rijn, H. (2018). Deploying a model-based adaptive fact-learning system in a university course. In *In proceedings of the 16th international conference on cognitive modeling* (p. 136-137). Madison, WI.

van Rij, J., Vaci, N., Wurm, L. H., & Feldman, L. B. (2020). Alternative quantitative methods in psycholinguistics: Implications for theory and design. *Word Knowledge and Word Usage*, *83*.

Van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *In proceedings of the 9th international conference of cognitive modeling* (p. 110-115). Manchester, United Kingdom.

Van Woudenberg, M. (2013). Optimal word pair learning in the short term: Using an activation based spacing model. *Unpublished master's thesis, University of Groningen*.

Verhoeven, F. M., & Newell, K. M. (2018). Unifying practice schedules in the timescales of motor learning and performance. *Human movement science*, *59*. doi: 10.1016/j.humov.2018.04.004

Wamsley, E. J. (2019). Memory consolidation during waking rest. *Trends in cognitive sciences*, *23*(3), 171-173. doi: 10.1016/j.tics.2018.12.007

Wood, S. (2006). Generalized additive models: An introduction with r. chapman and hall/crc. *Texts Stat. Sci.*, *67*, 391.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3-36. doi: 10.1111/j.1467-9868.2010.00749.x

Appendix A

Quintile plot bioPsych

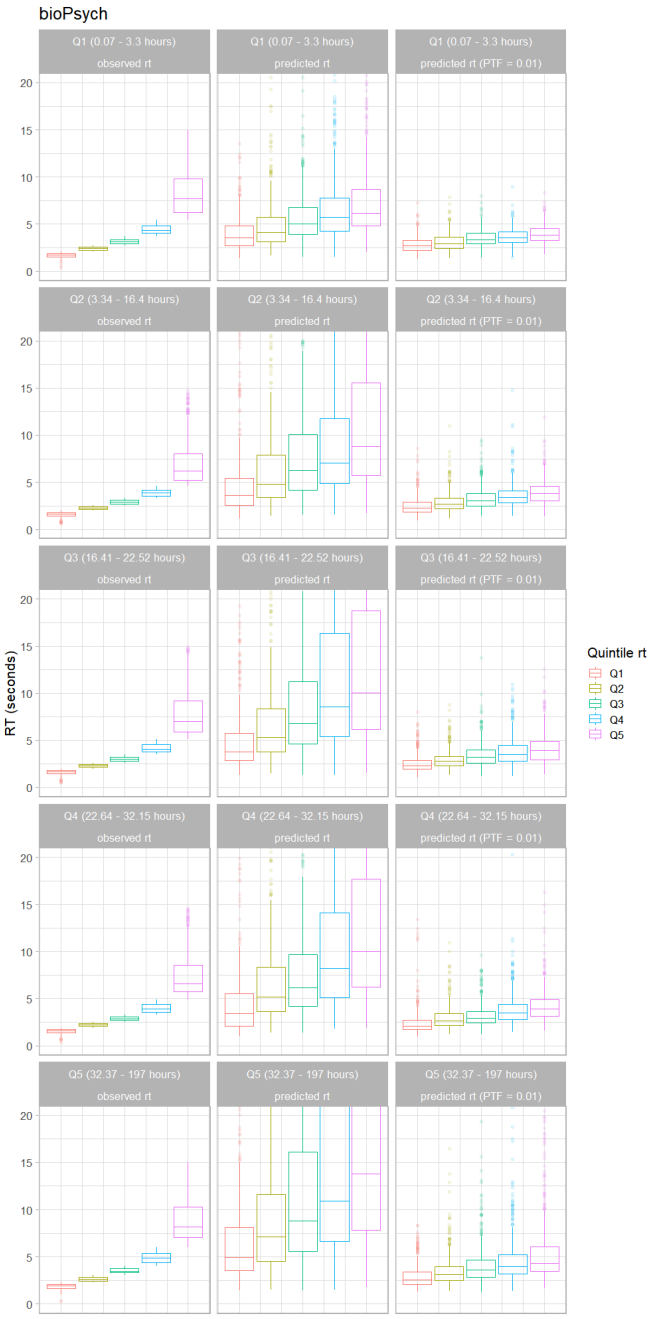

Figure A1. Quintile plot for response times in the bioPsych dataset grouped by intervals (rows) and by observed response times (columns). Boxplots of the observed response times (left), predicted response times in the clock condition (middle) and predicted response times in the psychological time condition (right) (PTF = 0.01) are shown. The whiskers of the boxplots are drawn as in prior plots.

Appendix B

Grid of PTFs over intervals between sessions with RMSE

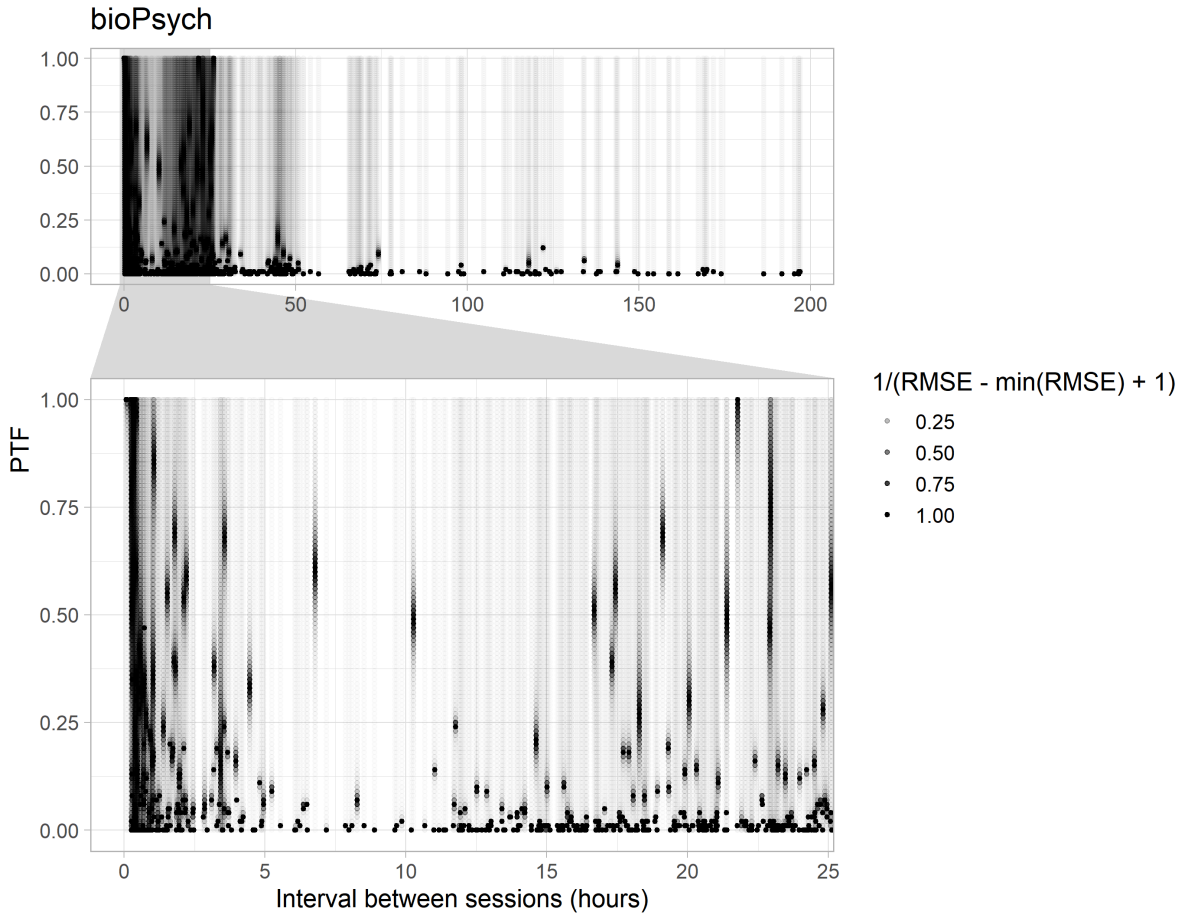

Figure B1. The calculated PTF values are shown over intervals between sessions in hours for the bioPsych dataset. The opacity of each point represents the transformed RMSE value. Opacity of 1 indicates that this PTF leads to the lowest RMSE value. Higher RMSE values are indicated by a lower opacity. The zoomed in plot shows the interval from 0 to 25 hours.

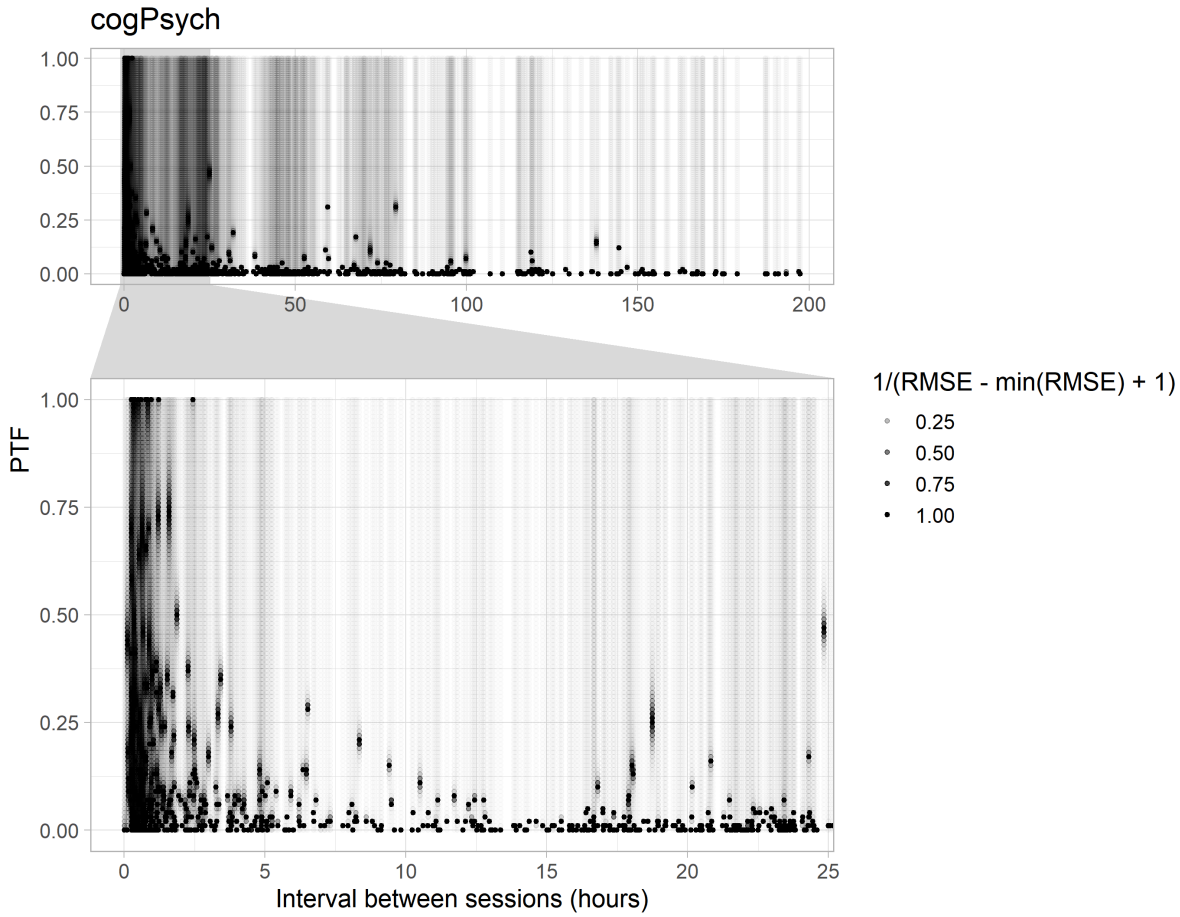

Figure B2. The calculated PTF values are shown over intervals between sessions in hours for the cogPsych dataset. The opacity of each point represents the transformed RMSE value. Opacity of 1 indicates that this PTF leads to the lowest RMSE value. Higher RMSE values are indicated by a lower opacity. The zoomed in plot shows the interval from 0 to 25 hours.

Appendix C

GAM

Following GAM functions were used in this thesis. The first model predicts the PTF_{Best} with the interval between sessions as a main effect.

$$\text{bam}(PTF_{best} \sim \text{s(interval, k = 10)}) \tag{7}$$

The second model predicts the PTF_{Best} with the interval between sessions and night between sessions as a main effects and the interaction of both.

$$\text{bam}(PTF_{best} \sim \text{night + s(interval, k = 10) + s(interval, by = night, k = 20)}) \tag{8}$$

Appendix D

AIC

This shows the equation for the AIC value used in the R package.

$$AIC = n + n \cdot log(2 \cdot \pi) + n \cdot (log(RSS/n)) + 2 \cdot (k + 1)) \tag{9}$$