

Pupil Dilation as an Alternative Signal for Adaptive Memory Modelling

Baran Akkanat
University of Groningen
The Netherlands

Abstract

Adaptive learning systems such as SlimStampen estimate how quickly a learner forgets an item from response time and accuracy. This study asks whether task-evoked pupil dilation can replace response time in that estimate. Two versions of the same model were compared: both used accuracy, but one updated the rate of forgetting from response time and the other from the pupil response measured in a three-second, response-free retrieval window. Eleven participants completed both conditions in a counterbalanced within-subject design, 90 trials per session, each session followed by a four-alternative knowledge test.

On items that were actually presented during learning, test accuracy was 93.5% in the pupil condition and 85.3% in the response-time condition (+8.2 percentage points, $d_z = 0.62$, exact permutation $p = .086$); for the nine participants whose two sets were tested after matched delays it was +12.0 points ($p = .031$). Replaying the sessions through the model showed that response time rose as memory strength fell ($\rho = +0.25$) while the pupil response did not ($\rho = +0.002$), and that the two models produced almost the same schedule. Pupil dilation can therefore drive an adaptive schedule without any cost in learning outcome, but we found no evidence that it estimated forgetting; the extra retrieval practice created by the pupil condition's viewing window is a more plausible account of the difference.

Keywords

adaptive learning, pupil dilation, memory modelling, spaced repetition, reaction time, eye tracking, SlimStampen

1 Introduction

Adaptive learning systems schedule repetitions so that an item is restudied shortly before it is forgotten. To do this they need an estimate of the learner's current memory strength.

SlimStampen, an adaptive fact-learning model derived from the ACT-R account of practice and forgetting [6, 10], estimates an item-specific rate of forgetting from the learner's behaviour: mainly how long the learner takes to respond, together with whether the response is correct. Rates estimated this way are stable within a learner and differ across materials [8].

Response time is a convenient but indirect measure. It covers the whole path from retrieval to a button press and is affected by motor speed and response caution as well as by memory. Pupil dilation is a second index of retrieval effort that does not depend on how fast a button is pressed: the pupil dilates with cognitive load [2, 3], and the dilation during retrieval is larger when the memory trace is weaker [5, 9].

This study asks whether that relation is strong enough at the level of a single trial to replace response time in the model.

1.1 Research Question

Can task-evoked pupil dilation be used as an alternative to reaction time for adapting a memory model, and how does learning performance differ between an RT-based and a pupil-based version of SlimStampen?

Three sub-questions follow: does the pupil-driven schedule lead to worse learning; does the pupil response carry information about memory strength on a single trial; and do the two signals make the model schedule practice differently.

1.2 Scope of This Paper

The experiment was designed and run by a five-person project group (Baran Akkanat, Jan Bethe, Javi Garrido Jimenez, Ania Strozyk and Moniek Schulting). The stimulus design, the experiment software and the data collection were group work; my own contribution was to the implementation of the experiment software and to running experimental sessions. This paper is an individual submission: the analyses, their interpretation and the text are my own.

2 Background

2.1 Adaptive Memory Modelling

SlimStampen estimates the activation $m_i(t)$ of item i at time t from the item's previous encounters,

$$m_i(t) = \ln \sum_{j \in E_i(t)} \left(\frac{t - t_j}{1000} \right)^{-d_j}, \quad (1)$$

where the decay of an encounter is itself activation-dependent,

$$d_j = c e^{m_j} + \alpha_i, \quad c = 0.25. \quad (2)$$

The item-specific parameter α_i is the rate of forgetting: a higher value means faster forgetting. The scheduler repeats the item whose predicted activation has fallen below a threshold τ , and introduces a new item when no studied item is due.

From the third encounter with an item onwards, the model predicts the response time implied by the current activation, compares it with the observed response time, and searches for the value of α_i that minimises the prediction error over the item's five most recent encounters.

2.2 Pupil Dilation and Memory

Van Rijn et al. [9] found that pupil dilation during retrieval decreased with repetition and was smaller for recently encountered items, which suggests that the dilation reflects retrieval effort. Pupil size also responds to luminance, gaze position, arousal and fatigue, so a usable measure needs a fixed display, a baseline taken immediately before the event of interest, and explicit handling of blinks [3].

3 Method

3.1 Participants and Design

Eleven participants completed both conditions (22 sessions), on 16 and 19 October 2026 at the University of Groningen. Participation was voluntary and informed consent was obtained.

The design was within-subject. Each participant learned one 15-item set under the RT model and the other set under the pupil model, in two sessions of 90 trials. Order and the assignment of sets to conditions were counterbalanced: the pupil condition came first for 6 participants and second for 5, and used set A for 5 and set B for 6.

For the first two participants both knowledge tests were given at the end of the second learning session; from the third participant on, each set was tested immediately after its own session so that the two conditions had the same retention interval. Results are therefore reported both for all eleven participants and for the nine with matched intervals.

Four earlier sessions in which the rate of forgetting was frozen for parameter estimation, and three aborted sessions, are excluded from all analyses.

3.2 Materials and Procedure

The materials were 30 map outlines in two sets of 15. Within a set, the outlines form five triplets of three visually similar shapes, and each outline is paired with a pronounceable pseudoword such as *Velkira*. Real place names were replaced to remove prior knowledge, and each outline received its own stem so that similar shapes did not also carry similar names. Every outline was drawn in the same light grey on the same mid-grey background as the screen and scaled to the same filled area, so that each image carried the same amount of light and the pupil's light reflex could not differ between items; the fixation disc had that same area and colour, so the change from fixation to outline involved no change in screen luminance. The four answer alternatives of a trial were always names from the participant's own set: the correct name, members of the item's own triplet and one name from the nearest other triplet, arranged so that every alternative was correct about one time in four.

A trial began with a fixation disc for 3 s, of which the last 250 ms served as the pupil baseline. In the pupil condition the outline then appeared alone for 3 s, during which no response was possible; in the RT condition the outline and the alternatives appeared together. Responses were given on a MilliKey SR-5 button box (four alternatives plus a middle button for "don't know"), followed by feedback of 800 ms after a correct answer and 3 s after an error. After each learning session the participant took a four-alternative test of all 15 items of that set, without feedback.

The experiment was implemented in OpenSesame [4] with the eye tracker controlled through PyGaze [1]. Pupil data were recorded with an SR Research EyeLink Portable Duo (host software EyeLink CL v6.54) in remote mode on a desktop mount, with the participant's head stabilised in a headrest, recording both eyes at 500 Hz after a 13-point calibration, with the tracker set to report pupil area; the online measure used the pupil size of one eye as returned by PyGaze. The experiment ran fullscreen at 1920 × 1080 on a 27-inch monitor (Philips 275B1, native 2560 × 1440, 60 Hz) at a viewing distance of about 75 cm, where the outline image subtended about 8.5° and the

filled shape about 3.5°, with the answer options between 9° and 11° from the centre. During the retrieval window the experiment read the pupil size from the tracker every 2 ms, which matches that sampling rate. The tracker ran in the pupil sessions only; the RT sessions used the same display, headrest and button box without eye tracking.

3.3 Pupil Measure

Following van Rijn et al. [9], the task-evoked response of a trial was the difference between the largest dilation and the largest constriction in the 3 s window after outline onset, as a percentage of the baseline. Samples were binned, gaps were interpolated and samples close to a blink were discarded; a trial counted as invalid when fewer than 70% of the samples in the baseline or the window survived. On a correct trial with an invalid measurement the pupil model made no update: no value was imputed. Because the tracker reported pupil area rather than diameter, a given dilation is about twice as large in percent here as in the diameter-based values usually reported.

3.4 The Two Models

Both conditions used the same scheduler, equations, estimation procedure and bounds (α starting at 0.3 and bounded to [0.1, 0.6], $\tau = -0.8$, 15 s look-ahead); the model code was identical. They differ only in the observation that is compared with the model's prediction.

The RT model uses the response time measured from the onset of the alternatives, so that the forced viewing window is never part of it, and compares it with

$$\widehat{RT}(m) = F e^{-m} \cdot 1000 + b, \quad F = 1.0, \quad b = 1000 \text{ ms}, \quad (3)$$

where b stands for the non-retrieval part of a multiple-choice response.

The pupil model replaces both quantities. It predicts a pupil deviation

$$\widehat{P}(m) = \beta (e^{-m} - e^{-\tau}), \quad (4)$$

which is zero for an item at the forgetting threshold and positive for a weaker memory, and compares it with the trial's pupil response relative to the participant's own average over the last ten valid repeat trials. In both models an incorrect answer counts as the strongest evidence of forgetting. The constants $\beta = 2.0$ and the error value of 10 percentage points were taken from the literature; they were not fitted to data.

3.5 Analysis

The primary outcome was accuracy on the knowledge test, restricted to items the scheduler had actually presented. With eleven participants, exact non-parametric tests are reported rather than mixed-effects models: an exact sign-flip permutation test over all 2^{11} assignments of the paired differences, with the Wilcoxon signed-rank test as a check, and Cohen's d_z as the effect size. The item-level analysis permutes the condition label within each participant. Activation per trial was recovered by replaying each session through the experiment's own model code.

4 Results

4.1 Data Quality

All 22 sessions ran to their full 90 trials. The scheduler introduced 13.7 of the 15 items on average in the RT condition and 13.3 in the pupil condition; because it introduces a new item only when no studied item is due, between 9 and 15 items were introduced per session, while the test asks all 15. Items that were never presented are therefore excluded from the primary analysis. A usable pupil response was obtained on 85% of trials on average, with one session as low as 52%.

A software fault in the version used meant that, after the first error of a session, correct answers also received the long feedback. It affected 66 of 90 trials in the RT sessions and 62 of 90 in the pupil sessions, a difference that was not reliable ($p = .635$), so it is balanced across the comparison.

4.2 Knowledge Test

Table 1: Knowledge test accuracy, with standard deviations, for all participants and for the nine whose two sets were tested after matched delays. Exact permutation tests over participants.

	RT	Pupil	Difference
<i>All participants (n = 11)</i>			
All 15 items (%)	83.0 (18.9)	90.9 (8.6)	+7.9, $p = .148$
Presented items (%)	85.3 (15.7)	93.5 (6.1)	+8.2, $p = .086$
<i>Matched retention (n = 9)</i>			
All 15 items (%)	79.3 (19.0)	91.1 (9.4)	+11.9, $p = .062$
Presented items (%)	82.0 (15.5)	94.0 (6.6)	+12.0, $p = .031$

On presented items the pupil condition was better by 8.2 percentage points (95% CI +0.4 to +16.1, $d_z = 0.62$, Wilcoxon $p = .141$), better for 6 of the 11 participants, equal for 3 and worse for 2 (Figure 1). The item-level analysis over all 297 presented items gives the same estimate and the same p value.

The only participants whose two conditions had unequal retention intervals are the first two, who were tested on both sets at the end, and in both the pupil set was the one tested later (1152 s and 1238 s against 401 s and 65 s). For the other nine the intervals are matched (mean 79 s against 72 s) and the difference is larger: +12.0 points (95% CI +4.5 to +19.5, $d_z = 1.04$, permutation and Wilcoxon $p = .031$), with no participant worse under the pupil model. The advantage is therefore not an effect of a shorter retention interval. The one participant whose pupil session fell below 60% valid trials is also one of those two; dropping that session from the full set as well gives +10.1 points ($p = .062$).

4.3 Learning and Model Behaviour

Accuracy on repeat trials was 86.0% (RT) and 91.0% (pupil), a difference that was not reliable ($p = .286$). The response time measured from the onset of the alternatives was shorter in the pupil condition (median 1795 ms against 2723 ms, $p = .002$), which follows from the design: the outline has already been visible for three seconds

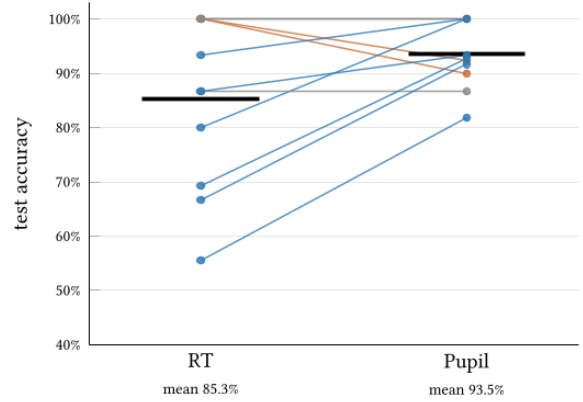


Figure 1: Knowledge test accuracy on presented items. Each line is one participant; blue rises with the pupil model, red with the RT model. The thick bars are the condition means.

when the alternatives appear. The same window made the pupil sessions about three minutes longer.

The models updated the rate of forgetting a different number of times: 61.2 of 90 trials in the RT condition against 52.8 in the pupil condition ($p = .001$), because the pupil model skips trials with unusable pupil data and needs a within-participant reference before it can use the signal. The resulting estimates were similar (mean α per session 0.301 against 0.321, $p = .388$; 0.29 and 0.32 averaged over items, the values used in Figure 3), and so were the schedules: median lag between two presentations of an item 60 against 65 s, and a similar spread of presentations across items.

4.4 Did the Signals Track Memory Strength?

Table 2: Correlation of each updating signal with e^{-m} , which increases as memory weakens. Correct repeat trials.

	RT model	Pupil model
Trials	720	653
Pooled ρ	+0.254	+0.002
Median per-participant ρ	+0.059	-0.043
In the predicted direction	7 / 11	5 / 11
Weaker half of trials	2987 ms	14.2%
Stronger half of trials	2433 ms	14.1%

Response time behaved as its model assumes: it was about 550 ms longer on the weaker half of trials. The pupil response did not vary with memory strength at all (Table 2). Activation predicted the answer in the pupil condition, from 83% in the weakest band to 100% in the strongest; in the RT condition the relation was irregular, with few trials in the extreme bands.

Figure 2 shows the same point from the data side. Accuracy plotted against the time since an item was last shown is flat, and in the RT condition it even rises. This is not an absence of forgetting but a consequence of adaptive scheduling: the model chooses the lag and repeats an item sooner exactly when it predicts the item is

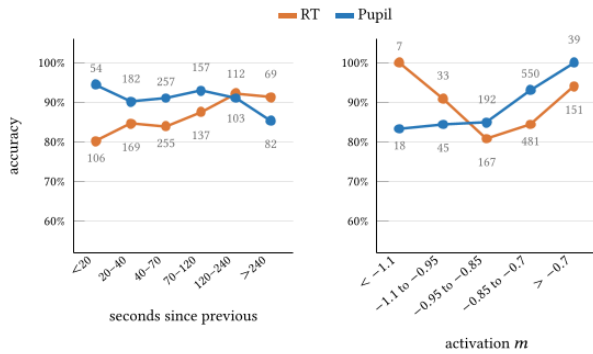


Figure 2: Accuracy against the time since the previous presentation (left) and against the activation at the moment of the trial (right). The small numbers are trial counts.

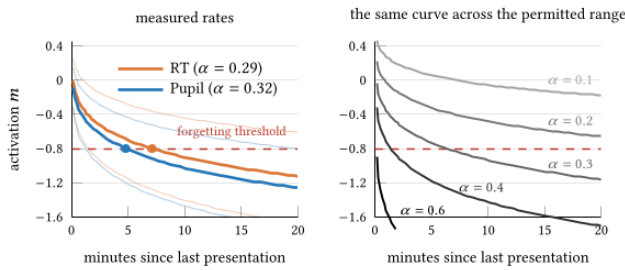


Figure 3: Forgetting curves implied by the estimated rates, computed with the model's own equations for an item practised 5 times at the median spacing of 60 s. Left: the mean rate of each condition, with faint lines at ± 1 standard deviation across items and a dot where the item falls below the threshold. Right: the same curve for rates across the range the model permits.

weak, so short lags contain the difficult items. Flattening this curve is what an adaptive scheduler is for.

4.5 What the Estimated Rates Predict

Figure 3 translates the estimated rates into the curve they imply, using the model's own equations for an item practised five times at the median spacing of 60 s. The RT model's mean rate brings the item to the threshold after 7.1 min and the pupil model's after 4.8 min. The difference between the conditions is small next to the range the model allows, where $\alpha = 0.1$ keeps an item above threshold for more than twenty minutes and $\alpha = 0.6$ brings it below within one, and it is also smaller than the spread between items inside a condition.

5 Discussion

Three results have to be read together. The pupil-driven schedule produced at least as good a test result as the response-time-driven schedule, and numerically better by about eight percentage points. The pupil response, as measured here, did not track memory

strength on a single trial, while response time did. And the two models scheduled practice almost identically.

Taken together these rule out the most attractive interpretation. The pupil condition cannot have done better because it estimated forgetting better: its estimates were unrelated to activation, and the schedule it produced was nearly the same as the one the RT model produced. The most plausible remaining explanation is the retrieval window itself. In the pupil condition every trial begins with three seconds in which the participant looks at the outline and tries to retrieve its name without being able to answer; over 90 trials this is roughly four and a half minutes of additional retrieval practice per session, and retrieval attempts before feedback are known to improve retention [7]. The design confounds the updating signal with that extra practice.

Why the pupil signal was flat is a separate question, and the answer is probably not that pupillometry cannot work. The task was easy (86–91% correct during learning), so most retrievals were not effortful, and a measure of retrieval effort shows little when there is little effort to measure. A usable response was obtained on only 85% of trials, the mapping from activation to expected dilation was never fitted to data, and a single trial's response is noisy, whereas the published demonstrations average over many trials.

For a practical system the conclusion is mixed. The pipeline works: an eye tracker can drive a repetition schedule online for a full session, and doing so cost nothing in learning outcome. But on this evidence the pupil signal added no information the model could use, while adding calibration, hardware, data loss and about 20% to the session length. A response-free retrieval window may itself be a useful intervention, independent of any pupil measurement.

6 Limitations

- Eleven participants; the primary comparison has a 95% interval of about ± 8 percentage points, so only large effects are detectable.
- Test accuracy on presented items was 82–94% and learning accuracy 86–91%, which leaves little room for either model to show an advantage.
- The pupil condition provides three seconds of extra retrieval practice per trial, which the design cannot separate from the updating signal. Only the RT condition asked for a fast answer, so the instruction differed as well as the timing.
- The feedback fault described above gave every participant more study time than the protocol specified, although it was balanced across conditions.
- The constants of both models were taken from the literature rather than fitted.
- Four alternatives drawn from a closed triplet allow answers to be found by elimination, and recognition differs from free recall.

7 Conclusion

Pupil dilation can serve as the input signal of an adaptive memory model: eleven full sessions ran with the schedule driven by live pupil measurements, and learning outcomes were at least as good as

with response-time scheduling, numerically better by about eight percentage points on the knowledge test.

There is, however, no evidence that the pupil signal estimated forgetting. It did not co-vary with the model's own measure of memory strength, the schedules it produced were almost the same as those of the response-time model, and the difference in test performance is more plausibly explained by the extra retrieval practice that the pupil condition provides. The defensible conclusion is that pupil-driven scheduling cost nothing, not that it gained anything.

Three steps would settle what this study could not: run the response-time condition with the same three-second window, so that only the updating signal differs; estimate the pupil parameters on data in which memory strength varies widely, after first checking that the pupil response and activation are related at all; and make the task and the test hard enough that both the outcome measure and the estimated rates have variance to show.

Acknowledgements

I thank Jan Bethe, Javi Garrido Jimenez, Ania Strozyk and Moniek Schulting for designing, building and running the study with me, and the instructors and laboratory staff of the University of Groningen for their support with the eye-tracking setup.

References

- [1] Edwin S. Dalmaijer, Sebastiaan Mathôt, and Stefan Van der Stigchel. 2014. PyGaze: An Open-Source, Cross-Platform Toolbox for Minimal-Effort Programming of Eyetracking Experiments. *Behavior Research Methods* 46, 4 (2014), 913–921.
- [2] Daniel Kahneman and Jackson Beatty. 1966. Pupil Diameter and Load on Memory. *Science* 154, 3756 (1966), 1583–1585.
- [3] Sebastiaan Mathôt. 2018. Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition* 1, 1 (2018), 16.
- [4] Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes. 2012. OpenSesame: An Open-Source, Graphical Experiment Builder for the Social Sciences. *Behavior Research Methods* 44, 2 (2012), 314–324.
- [5] Megan H. Pappas, Stephen D. Goldinger, and Michael C. Hout. 2012. Memory Strength and Specificity Revealed by Pupillometry. *International Journal of Psychophysiology* 83, 1 (2012), 56–64.
- [6] Philip I. Pavlik and John R. Anderson. 2005. Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. *Cognitive Science* 29, 4 (2005), 559–586.
- [7] Henry L. Roediger and Jeffrey D. Karpicke. 2006. Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science* 17, 3 (2006), 249–255.
- [8] Florian Sense, Friederike Behrens, Rob R. Meijer, and Hedderik van Rijn. 2016. An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science* 8, 1 (2016), 305–321.
- [9] Hedderik van Rijn, Jelle R. Dalenberg, Jelmer P. Borst, and Simone A. Sprenger. 2012. Pupil Dilation Co-Varies with Memory Strength of Individual Traces in a Delayed Response Paired-Associate Task. *PLoS ONE* 7, 12 (2012), e51134.
- [10] Hedderik van Rijn, Leendert van Maanen, and Marnix van Woudenberg. 2009. Passing the Test: Improving Learning Gains by Balancing Spacing and Testing Effects. In *Proceedings of the 9th International Conference on Cognitive Modeling*. University of Manchester, Manchester, UK, 110–115.

A Model and Experiment Parameters

Table 3: Parameters used in the study.

Parameter	Value
Initial rate of forgetting α	0.3 (bounds 0.1–0.6)
Forgetting threshold τ	−0.8
Look-ahead	15 s
Decay constant c	0.25
F , base time b (RT model)	1.0, 1000 ms
β , error value (pupil model)	2.0, 10 points
Reference window (pupil model)	last 10 valid repeat trials
Pupil baseline / analysis window	250 ms / 3 s
Minimum valid sample fraction	0.70
Inter-trial interval	3 s
Feedback, correct / incorrect	800 ms / 3 s
Display	1920 × 1080 fullscreen, 60 Hz
Viewing distance	about 75 cm (estimated)
Trials per session	90
Items per condition	15
Answer alternatives	4 (+ “don’t know”)