The Joy of Learning

Bachelor's Project Thesis

Jesse Stoffels, s2919079, j.a.stoffels@student.rug.nl,
Supervisors: F.Cnossen, J.Georgiadis, A.Van Gaalen

1 Introduction

Mental arousal comes in varying degrees of intensity, ranging from apathy at the lower end to a panic attack at the upper end of the spectrum. Somewhere between these extremes lies an optimal state of attentiveness, resulting in complete engagement and focus, causing people to lose track of time and unrelated stimuli due to their complete immersion in their task. This state is known in psychology as flow. Flow as a psychological construct was named as such by Mihaly Csikszentmihalyi in the book "Beyond boredom and anxiety" (Csikszentmihalyi, 1975). It describes flow as a state of optimal engagement. Earlier mentions of a similar concept do exist, though Csikszentimihalyi uses the term 'play' in these instances (Csikszentmihalyi and Bennet, 1971), after the work "Homo Ludens" by Dutch cultural historian /cultural scientist Johan Huizinga (Huizinga, 1938). In all instances the sensation was considered positive and linked to engagement in the activity.

Education is one particular field that could benefit greatly from this sensation since a positive experience and engagement are often absent when forced to learn. If flow could be generated while learning education could become an innately rewarding fun experience for the students instead of a chore, as is regularly the case.

One of the core requirements to reach a flow state is matching challenge and ability (Csikszentmihalyi and Nakamura, 2009). To see if this is achievable in learning we must first specify what the challenge in learning is. This varies from task to task, such as application or spotting parallels but for the purpose of this paper we will focus on recollection. In recollection the difficulty is retrieving the information from memory, either long term or working memory. The goal of studying is for the information to end up in long term memory so we will primarily look at that. A general equation used to represent a facts strength in memory is the ACT-R equation:

$$A_i = \sum_{j=0}^{n} (t - t_j)^{-d_j}$$

A_i represents the facts strength in memory or the activation of the fact, t is the point in time of said strength, t_j is the time of the presentation of the fact and d_j is a decay factor. The lower the activation, the harder the recollection. Should the activation fall below a certain threshold, the fact can no longer be (reliably) retrieved.

With the level of challenge not only defined but quantified we can attempt to match it to the ability of a student. By presenting facts as they are about to be forgotten, students are presented with the greatest amount of mental effort that they can successfully handle. This seems to fulfil the matching of challenge and ability required for flow. There is a spacing algorithm known as Slim Stampen (RL) that tracks the activation of facts and attempts to present them at that point in time. It uses the aforementioned equation and makes a dynamic estimate of the decay factor, tweaking it continuously to ensure that the facts are presented at the right time.

To test that the RUGged algorithm (Van Rijn, van Maanen, and van Woudenberg, 2009), and with it learning, can induce flow, it was tested against a more traditional ordering method: flashcards (FS). The chosen variant has a few facts that are tested in a loop, with every correct fact removed from the pool until it is empty. This method clusters the remaining facts more and more, increasing the chance that recollection comes from working memory and worsening long term retention. The level of challenge varies widely due to the changing time between repeated facts, so no flow is expected. Several flow measurements will be taken while learning with both order algorithms and this should answer the central question: Can flow be induced through learning?

To this end an experiment was conducted where participants had to learn two different sets of anatomical elements, one ordered according to the flashcards algorithm and one ordered according to Slim Stampen algorithm. They played a round of Tetris to clear their working memory after both sets and then they were tested on how well they learned. The various flow measures were compared between the two learning phases. Since flow is hard to compare between subjects a within-subject design was chosen.

2 Method

2.1 Participants

There were 35 participants. All participants were medical students students in the age range 19-27 (mean = 21, sd = 1.76). They were recruited through advertisement during their lectures. Though Anatomy is part of their curriculum, they had no prior experience with the particular elements in the experiment. 20 of the participants had prior experience with a Classical language (Greek or Latin) in high-school, which might affect their recall since the anatomical elements are named in Latin. 25 of the participants were native Dutch, the rest spoke various languages from French to Arabic. All were proficient enough in English to comprehend the instructions given to them over the course of the experiment. All participants were given 10 euros for their participation.

2.2 Measures

Determining if a state of flow was induced while learning requires 2 observations, namely whether or not learning occurred and if an enhanced state of flow was present during this.

The primary means of obtaining a measure for the state of flow in participants was the Flow Short Scale. The Flow Short Scale (FKS, Rheinberg, Vollmeyer, and Engeser (2003)) was a 16 question questionnaire designed to measure mental engagement and challenge. Responses were on a range from 1 to 7 with increasing scores indicating higher presence of flow. The tests is verified reliable reliable with a Chronbachs α of 0.90.

Heart-rate data was also measured in the form of electric potential over time, later to be processed into other useful metrics. An EMG was also taken for two muscles, also in the form of electric potential over time.

2.3 Materials

The experiment was run in an OpenSesame (Mathôt, Schreij, and Theeuwes, 2012) environment on a laptop. The anatomical elements the participants had to learn were taken at random from a list of 34. The Slim Stampen version and the Flashcards version were identical except for the elements and their the ordering: white background, black letters and a color picture of the skull with a green circle around the particular element that had to be identified. In all instances the background was white and the letters of the words were black, except the feedback which was green or red in case of a correct or incorrect answer respectively. Font size was 18.

A game of Tetris was also used between the learning phase and the testing phase. It was used primarily to clear the working memory. Since the game is traditionally enjoyable, Flow measures were also taken while they were playing.

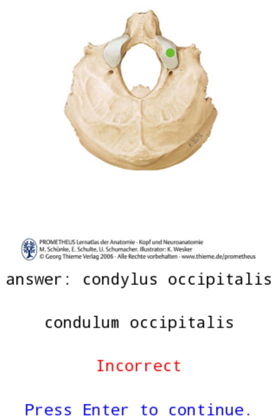

Figure 2.1: An example of how the elements were presented during the learning phases

2.3.1 Flashcards

The Flashcards model first showed the first series of 5 elements to be learned alongside their correct Latin names. These elements would then be shown, one by one, without their Latin names and the participant had to fill in the correct one. After the participant answered, they were given feedback. A correct answer would remove the element from the list while an incorrect answer meant it would be appear again. When the last element of those 5 had been correctly named, the process repeated with the second 5 elements to be learned. Single elements had no designated time limit but participants were informed about the overall time limit of 10 minutes.

2.3.2 Slim Stampen

The Slim Stampen model first showed the first element to be learned alongside its correct Latin name. The rest of the ordering was determined by the Slim Stampen algorithm. The algorithm keeps track of how often and when each element was presented and how well and fast it was answered. It uses all this information to estimate how present every element is in the participants memory. This 'presence in memory' is called the 'activation' of the element. If the estimated activation of an element dropped below a certain threshold, it would be show again as the next element. At that moment the participant should barely remember the element, making it the most effective moment to present it. If the estimated activation of all elements was above the designated threshold, a new element would be shown. After every answer the participant would receive feedback on the correctness of their answer as well as the intended answer. As with the Flashcards model there was no time limit on individual elements but participants were informed about the overall time limit.

2.4 Data Collection

Test data was extracted from the OpenSesame logs. The program logged a wide variety of metrics though only the number of presented elements during training and the number of remembered elements during testing were used for this study.

The heart-rate data was obtained using a Cortrium C3 wearable heart-rate monitor placed over the sternum. It ran over the course of the experiment and was not disturbed between placement and removal.

Five electodes were placed, 2 on the Zygomaticus Major, 2 on the Corrugator Supercilii and an earthing wire placed on the C7 vertebrae. The electrodes were placed on the right side of the face and were bundled and secured using painters tape. The wires and electrodes were kept out of the participants field of vision as much as possible and it was verbally confirmed they were no distraction. The participants were also given instructions not to touch or otherwise disturb the wires during the experiment to prevent erroneous signals.

2.5 Design

The experiment had a within-subject design. All participants did both conditions in one session of approximately one hour. No true breaks were scheduled though subjects could relax at certain moments of the experiment. During both conditions subjects were given 15 anatomical elements to learn chosen at random from a list of 40. The starting condition was alternated to counterbalance any effect ordering might have had.

2.6 Procedure

Before the start of the experiment participants had to hand in a signed informed consent form. One was sent over mail and most participants brought read and signed forms to the experiments. The remainder was given the form at the location. First the participant was instructed on how to apply the Cortium heart-rate monitor and it was confirmed that it was placed correctly. Afterwards the EMG equipment was attached. First the earthing electrode was attached to the neck and they were instructed to turn to face the experimenter for the other electrodes. After the placement of both pairs of electrodes their placement was checked. The electrodes had to give a clearer signal for the intended muscle than for other close or prominent muscles to avoid false readings. This was done by requesting facial expressions and checking the EMG activation in real time. Any visible wires and loose electrodes were secured with painters tape and the wires were bundled and led past the participants ear to reduce the chance of electrodes falling off or distracting the participant.

After all the equipment had been attached the participant began the experiment. This consisted of two largely equal halves. Both began with a 5 minute calibration period for the heart rate. The participant was instructed to relax and not move excessively. Then the experimenter approached to give instructions about the learning phase. These included a summary of the task at hand and a note that any spelling errors or double spaces would invalidate an answer. Following a 10 minute learning phase the experimenter started the FKS questionnaire. This had no time limit, to prevent rushing the participant and compromising the results. The FKS was taken digitally and once the participant had finished, the experimenter came over and started the 5 minute game of Tetris for the subject. When the time was up a second FKS followed and after that the test phase started. During this phase all 15 words selected for the learning phase were shown, regardless of whether or not the participant had seen them, and no feedback was given after answers. The participant was instructed to ignore any words they hadn't seen yet and simply skip them. Once this had concluded the first time, the entire procedure was repeated for the other algorithm. After the second half the participant was debriefed.

2.7 Data Processing

The FKS data was averaged to result in a single flow measure for each individual.

The Heart-rate data was fed through pre-CAR (van Roon and Mulder, 2012) before being inspected and corrected manually to prevent false readings. The first and last minute of every measurement period was cropped due to timing inaccuracies. The interbeat interval data file produced by pre-CAR was then fed to Carspan (Mulder, Van Roon, and Schweizer, 1995). The 8 minutes of training data were divided into sections called blocks. The lengths were 2 block of 3 and 1 block of 2 minutes. The rest and Tetris sections were left as a signle block of 3 minutes. Carspan did a power spectrum analysis on every of the aforementioned blocks. The low frequency band was 0.02 - 0.06 Hz, the mid frequency band was 0.07 - 0.14 Hz. and the high frequency band was 0.15 - 0.40 Hz.

The EMG should have been passed through a low-frequency filter and a high-frequency filter before being conjugated with a previously taken high activation baseline. Since this baseline was not taken during the experiment, the EMG data was unusable and will be omitted.

2.8 Analysis

Asserting that learning occurred was done using the scores of the testing phase. Since they had no prior knowledge, all correct answers are assumed to have been learned. A pairwise t-test was used be used between the test scores of both algorithms and between the first and last trial.

To determine the state of flow of the participants during the different phases of the experiment we analysed Heart rate and FKS scores. The FKS data was analysed with a pairwise t-test between the different algorithms.

The different HR block were expressed as the percentage difference from the preceding resting period. That data was converted then to natural logarithms. A repeated measures ANOVA was performed to for all frequency bands as well as the Interbeat interval . All data sets were tested and corrected for sphericity using Mauchleys test of Sphericity. Below epsilon 0.75 Greenhouse-Geisser correction was used, above Huynh-Feldt was used.

3 Results

3.1 Test Scores

One participant was excluded due to an error in element presentation, which rendered their data unusable. The two sets of test scores were analysed through a paired t-test. The mean test score for the Slim Stampen algorithm (M (Mean) = 9.48, SD (Standard Deviation) = 3.355) was slightly higher than for the Flashcards algorithm (M = 8.79, SD = 4.189) though the difference was not statistically significant $t(32) = -.910$, $p = ,369$. The data showed the Flashcards method presented roughly 2-3 items more than the Slim Stampen method yet there was no significant increase in the amount of correct answers. To verify this a paired sample t-test was done on the number of presented items. This test showed that the amount of items seen during the training phases with the Flashcards method (M= 14.58, SD = 1.501) was indeed significantly higher than during the trails with the Slim Stampen method (M = 12.85, SD = 2.451), $t(32) = 3.477$, $p = < 0.001$. The accuracy was then calculated as the number of correct items over the number of seen items for each participant and training phase. Theses results were then subjected to another paired-sample t-test. This test showed that the accuracy for the training phases with the Slim Stampen method (M = .720, SD = .171) was significantly higher than for training phases with the Flashcards method (M = .592, SD = .272), t(32) = -2.80, p = < 0.001.

These tests were also performed comparing the fist and second testing phase, regardless of which algorithm was used. The average score for the first training phase (M = 7.85, SD = 3.759) was significantly lower than for the second training phase (M = 10.42, SD = 3.391), p(32) = -4.101, p = < 0.001. The average amount of seen items in the first training phase (M = 13.06, SD = 2.657) was significantly lower than in the second training phase (M = 14.36, SD = 1.365), t(32) = -2.433, p = 0.0208. The accuracy of first trails (M = .588, SD = .234) was, on average, significantly lower than than for second training phase (M = .724, SD = .217), t(32) = 3.03, p = < 0.001.

3.2 Flow Short Scale

Paired t-test showed no significant difference between the scores of tests taken after Flashcards blocks (M = 4.86, SD = .723) and those taken after Slim Stampen blocks (M = 4.96, SD = .746), t(32) = .838, p = .174. A different comparison did show a significant effect: FKS scores after the second block (M = 4.98, SD = .731) were, on average, slightly higher than FKS scores after the first block (M = 4.84, SD = .734), t(32) = -2.045, p = .049. This effect was absent in the tests following the rounds of Tetris. The tests after the first round (M = 4.73, SD .667) were not significantly different from the tests after second round (M = 4.60, SD = .674), t(13) = -1.391, p = .417.

3.3 Heartrate

Three participants were excluded from the heart rate analysis due to excessive noise. On all 3 frequency bands and the average Interbeat Intervals a repeated measures ANOVA was performed. After the Greenhouse-Geisser correction it was determined that the different blocks had a significant effect on the average Interbeat Interval (F(2.957, 85.744) = 16.562, p = < 0.001). In learning blocks the average IBI is lower than in rest or tetris blocks. This would suggest that the subjects are less relaxed when learning. There is also an upward trend over time, suggesting subjects calmed down over time.

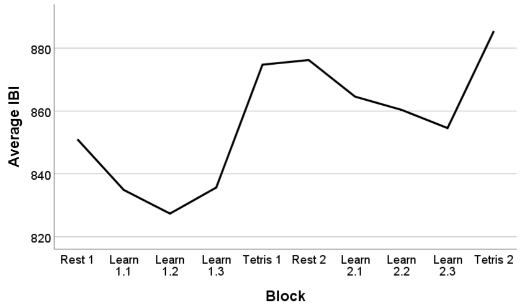

Figure 3.1: A graph of the average Interbeat Interval per block. The blocks are ordered over time.

Figure 3.2: A graph of the average power for all three frequency bands per block. The blocks are ordered over time.

After the Greenhouse-Geisser correction it was determined that the different blocks had no significant effect on the low frequency band power (F(4.066, 117.919) = .426, p = .792).

A similar approach was taken for the mid-frequency band. After the Greenhouse-Geisser correction it was determined that the different blocks had no significant effect on the mid frequency band power (F(4.287 , 124.323) = 1.851, p = .119).

The same was done for the high-frequency band. After the Greenhouse-Geisser correction it was determined that the different blocks had no significant effect on the high frequency band power (F(2.944, 85.362) = .535, p = .656).

4 Discussion

4.1 Summary

This study was set up to explore the possibilities of inducing flow through learning. Two learning algorithms were used, Slim Stampen and a Flashcard algorithm. The Slim Stampen algorithm was expected to significantly outperform the Flashcards algorithm, both in test scores and flow measurements. The results did show a higher retention percentage for the Slim Stampen method than for the Flashcards method though the flow measurements showed no significant difference between the two.

The average test scores were not significantly different for the two algorithms but the Flashcards algorithm showed, on average, more items than Slim Stampen. The average ratio of words seen to words learned was higher for Slim Stampen than for Flashcards.

The FKS questionaire showed no significant difference between Flashcards trails and Slim Stampen trails. This does not align with the hypothesis that Slim Stampen induces flow better than Flashcards. An interesting pattern that was observed was a higher average FKS score for the second trail compared to the first trail. Statistical tests indicated that this difference was indeed significant. This might be caused by a clearer expectation going into the second trail, increased confidence following the first test or them growing accustomed to the experiments conditions (the face full of sensors and wires) though the experiment wasn't setup to determine the exact cause of the increase. Acclimatisation as the primary cause of the flow increase is challenged by the FKS scores after the Tetris trails. Comparing these first to second showed no significant difference between the two, making it more likely that the flow increase was related to the learning trails rather than a general increase over time.

The other metric used for measuring flow was heart rate. Prior research has suggested correlation between flow and heart rate variability in the high and low frequency bands (Tozman, Magdas, MacDougall, and Vollmeyer (2015)). This research showed that when the task is excessively challenging the low and high frequency decrease and vice versa. The expectation was that the less consistent difficulty of the Flashcards algorithm would cause more variance of these variables between different phases of the trails when compared to the more consistent difficulty of the Slim Stampen algorithm. Lower values for the low frequency band and moderate levels for the high frequency band were also expected for the Slim Stampen, since this is what the research of Tozman et all showed was associated with higher levels of flow. The repeated measures ANOVA showed no effect of the different blocks on any of the frequency bands. This contradicted both hypotheses since both assumed different blocks having an effect on the data. The different blocks did have a significant effect on other physiological measurements, namely average heartrate and average interbeat interval, though these measurements have yet to be linked to psychological phenomena.

4.2 Limitations

The study had some setup issues that might have affected data. EMG data was collected for all test subjects but no baseline was taken, rendering all that data unusable.

Timing inaccuracies between the heart rate monitor and the start and end of the trails cost one minute on both sides of the data. This equated to about 20% of usable data. Marking the start of the different phases in the heart rate file or central timing mechanism for all aspects of the experiments would greatly reduce these data losses and might yield different results.

Another change that might prove useful is extending the training periods. At the current length a large portion of participants made it to the last cycle of items in the flashcards algorithm yet few completed it. This means that some items saw little to no repetition. During Slim Stampen trails, items saw more repetition but because of this most participants never saw all the items. Longer training periods would address both these issues and might make differences between the algorithms clearer.

4.3 Conclusion

In conclusion this study found that the Slim Stampen algorithm increased accuracy over the Flashcards algorithm during the tests but presented fewer words overall. However no evidence that either algorithm induced increased flow in the participants was found. While increasing engagement in learning might be possible, optimizing spacing does not appear as to be a viable method for doing so. At least as of yet.

References

M. Csikszentmihalyi. *Beyond boredom and anxiety.* Jossey-Bass Publishers, 1975.

M. Csikszentmihalyi and S. Bennet. *An exploratory model of play*, volume 73. 1971.

M. Csikszentmihalyi and J. Nakamura. Flow theory and research. In C. R. Snyder and S. J. Lopez, editors, *Oxford Handbook of Positive Psychology*, chapter 18, pages 195–206. Oxford University Press, Oxford, 2009.

J. Huizinga. *Homo Ludens: Proeve Ener Bepaling Van Het Spelelement Der Cultuur.* Wolters-Noordhoff, 1938.

S. Mathôt, D. Schreij, and J. Theeuwes. Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44:314–324, 2012.

LJM Mulder, AM Van Roon, and DA Schweizer. *CARSPAN: cardiovascular data analysis environment: user's manual.* iec ProGAMMA, 1995.

F Rheinberg, R. Vollmeyer, and S. Engeser. Die erfassung des flow-erlebens. In J. Stiensmeier-Pelster and F. Rheinberg, editors, *Diagnostik von Motivation und Selbstkonzept*, pages 261–279. Universitt Potsdam, Potsdamn, 2003.

T. Tozman, E. S. Magdas, H. G. MacDougall, and R. Vollmeyer. Understanding the psychophysiology of flow: A driving simulator experiment to investigate the relationship between flow and heart rate variability. *Computers in Human Behavior*, 52:408–418, 2015.

Hedderik Van Rijn, Leendert van Maanen, and Marnix van Woudenberg. Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th International Conference of Cognitive Modeling*, volume 2, pages 7–6, 2009.

A.M. van Roon and L.J.M. Mulder. *CARSPAN: cardiovascular data analysis environment: user's manual.* iec ProGAMMA, 2012.

A Item list

1. clivus
2. foramen magnu
3. gonion
4. lambda
5. mandibula
6. maxilla
7. os frontale
8. os nasale
9. os occipitale
10. os palatinum
11. os parietale
12. os petrosum
13. os sphenoidale
14. os zygomaticus
15. sutura sagittalis
16. vomer
17. canalis occipitalis
18. condylus occipitalis
19. crista galli
20. fissura orbitalis superior
21. foramen incisivum
22. foramen infraorbitale
23. foramen jugulare
24. foramen mandibulae
25. foramen mentale
26. foramen spinosum
27. os ethmoidale
28. os lacrimale
29. processus condylaris
30. processus coronoideus
31. processus mastoideus
32. processus styloideus
33. processus zygomaticus
34. sella turcica