Distracted in a Demanding Task: A Classification Study with Artificial Neural Networks

Stefan Huijser[1], Niels Taatgen, Marieke van Vugt

Institute of Artificial Intelligence, University of Groningen
Nijenborgh 9, 9747 AG Groningen, Netherlands
[1]s.huijser@rug.nl

Abstract. An important issue in cognitive science research is to know what your subjects are thinking about. In this paper, we trained multiple Artificial Neural Network (ANN) classifiers to predict whether subjects' thoughts were focused on the task (i.e., *on-task*) or if they were distracted (i.e., *distracted thought*), based on recorded eye-tracking features and task performance. Novel in this study is that we used data from a demanding spatial complex working memory task. The results of this study showed that we could classify on-task vs. distracted thought with an average of 60% accuracy. Task performance was found to be the strongest predictor of distracted thought. Eye-tracking features (e.g., pupil size, blink duration, fixation duration) were found to be much less predictive. Recent literature showed potential for eye-tracking features, but this study suggests that the nature of the task can greatly affect this potential. Rehearsal effort based on eye-movement behavior was found to be the most promising eye-tracking feature. Although speculative, we argue that eye-movement features are independent of the content of distracted thought and may therefore provide a more generic feature for classifying distracted thought.

Keywords: Distracted Thought, Mind-wandering, Demanding Task, Artificial Neural Networks.

1 Introduction

One important challenge cognitive scientists face in their research is to determine what someone is thinking about: Are my subjects doing the task, how are they solving it, have they wandered off? In particular the latter question is challenging. Knowing whether someone is distracted is difficult, as it is usually not accompanied with (outwardly) observable behavior. Researchers that study mind-wandering, a self-generated and task-unrelated thought process [1], have commonly solved this problem by using self-report. This method can give you a peak into the richness of thought. Although shown to produce valid results [2], it is hard to implement them correctly [3]. Furthermore, the discrete nature of the method leaves the researcher oblivious to the dynamics of the thought process. For these reasons, mind-wandering researchers have explored alternative methods to measure mind-wandering.

One candidate method is to infer the characteristics of distracted vs. non-distracted

thought from eye-tracking measurements. Distracted thought refers here to all thinking that is irrelevant to performing the task at its time of occurrence. Eye-tracking recordings are often conducted in conjunction with self-report, allowing researchers to track thought throughout the experiment. Examples of such research have identified that mind-wandering and inattentiveness (i.e., my mind is 'blank') are related to lower phasic and tonic pupil activity [4–6]. Studies with reading tasks have found that mind-wandering results in longer and more frequent fixations [7]. Also, it has been shown that increased and more prolonged blinking is correlated with mind-wandering [8]. Taken together, these studies show that distracted thought such as mind-wandering results in measurable patterns in eye-tracking features.

The successes in associating features from eye-tracking recordings to mind-wandering raised the question whether such features are also predictive of mind-wandering, or distracting thought in general, on a single-trial level. Recently, a study by Grandchamp and colleagues [8] explored this question by training a support vector machine classifier to predict distracted vs. non-distracted thinking with a collection of eye-tracking features. The researchers reached a classification accuracy of 80%. Noticeably, pupil size was found to be the most reliable predictor, with low average pupil size being indicative of distracted thought. In conclusion, this study showed that eye-tracking features, with pupil size in particular, could be used to classify distracted thinking in our subjects.

In this paper, we will also examine how eye-tracking related features are related to distracted thought, by training an artificial neural network classifier to predict if subjects were performing the task (i.e., *on-task thought,* OTT) or experienced *distracted thought* (DT). Novel in this study is that we will use data from a relatively demanding task, whereas previous classification studies only used simple task paradigms. In addition, we will consider both eye-tracking related features and trial-to-trial task performance in classifying on-task and distracted thought examples. Including both will allow us to compare the contribution of eye-tracking features, which can be measured continuously, with task performance measured only on a trial-to-trial basis.

2 Methods

The data set we use in this study was collected in a previous eye-tracking experiment in which subjects performed a working memory task and occasionally reported if their recent thought was on-task or distracted. We refer to Huijser, van Vugt, and Taatgen [9] for full details on the task, materials, and data acquisition.

2.1 Subjects

In total 38 individuals agreed to participate in the experiment. All subjects were native Dutch speakers and had normal or corrected-to-normal vision. We excluded the data of 5 subjects due to excessive data loss in the eye-tracking recordings. In addition, we removed the data of on extra subject, as this subject did not perform the task as required. This left us with 32 subjects for analysis and classification. All subjects

signed an informed consent prior to the start of the experiment and received a small monetary compensation when finished.

2.2 Task

The task performed by the subjects was a spatial complex working memory (SCWM) task. This task required subjects to memorize a sequence of targeted locations in a 4x4 grid, while also performing a processing subtask in between each presented location. On every trial, we first presented a storage target for 1 second. This allowed the subjects to encode the location of the target. Following this storage phase, a 4 seconds self-paced processing phase started, in which subjects made binary decisions (i.e., yes/no) on word stimuli presented in the same grid. The words moved every second to a random but different position to prevent visual rehearsal of the storage targets and to keep subjects engaged in the task[1]. Following the processing phase, the grid was emptied for 2 seconds, allowing subjects to rehearse or to potentially get distracted. The eye-tracking recordings during this 'blank' phase were used to classify examples as on-task thought or distracted thought (see section *Thought-probe* below).

Storage, processing, and blank phases were repeated a number of times equal to a span of three or four. The experiment included 96 trials, with half of the trials being of span 3 and the other half span 4.

2.3 Thought-probe

To determine whether subjects experienced on-task or distracted thought on a particular trial, we sampled recent thought content with thought-probes. Thought-probes are self-report questions aimed at assessing recent or current conscious experience, and were conducted on random but equally distributed moments in the experiment (i.e., half of the trials, resulting in 48 thought-probe trials).

In this experiment, we used an adapted version of a thought-probe question introduced by Stawarczyk, Majerus, Maj, Van der Linden, and D'Argembeau [10] and Unsworth & Robison [4]. The question was (translated from Dutch), 'What were you thinking about before you were prompted to answer?', with the following response options: (1) I tried to remember the location of the X's; (2) I was still thinking about the words from the decision task (= processing task); (3) I was evaluating aspects of the task (e.g., my performance, how long it takes, difficulty of the task); (4) I was distracted by my environment (sound/ temperature etc.) or by my physical state (hungry/thirsty); (5) I was mind-wandering/ I thought about task unrelated things, (6) I was not paying attention, but I did not think about anything specific. Response option 1 was labeled as *on-task thought* (OTT). The remaining options together were labeled as *distracted thought* (DT). Response option 3,4, and 6, were labeled as *task-related*

[1] The experiment of Huijser and colleagues [9] involved two conditions in the processing phase. These conditions used different word stimuli. As we were not interested in the conditions, we collapsed trials of both conditions in this study. Therefore, we do not discuss the two conditions here.

interference, external distraction, and *inattentiveness* respectively [see also 10]. We labeled option 2 as *mental elaboration* and option 5 as *mind-wandering* [see also 9].

2.4 Eye-tracking measurement

Equipment. An Eyelink 1000 eye-tracker system from SR Research was used to sample pupil size (in arbitrary units) and gaze position (X-, Y-coordinates in pixels) at a rate of 250 Hz. Blink events were automatically detected by the Eyelink software.

Preprocessing. Before training the classifier, we first performed a selection of preprocessing steps on the eye-tracking recordings. First of all, we discarded the pupil size and gaze position data for all automatically detected blinks including 100 ms before and after the blink events. In addition, we removed remaining artifacts by discarding sudden downward jumps (>200 units in pupil size, approx. 0.5 *SD*). Thereafter, we linearly interpolated the resulting missing data and subsequently downsampled the full data set to 100 Hz.

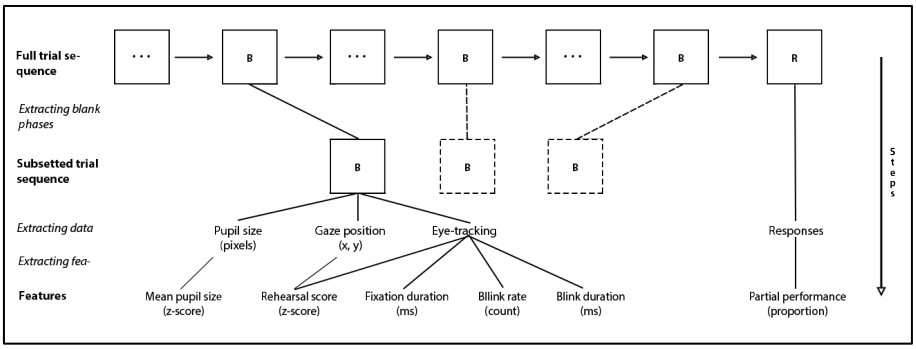

Fig. 1. Overview of the feature extraction procedure in a span 3 trial. To extract the features from the data set, we collected the responses and selected all the blank phases (i.e., thereby discarding the storage and processing phases). The recorded data in each blank phase and the collected response were subsequently used to compute the features. Square boxes with dots refer to the storage and processing phases. The square boxes with B refer to blank phases.

2.5 Classifier

We trained an artificial neural network (ANN) classifier [see e.g., 11] on the recorded eye-tracking and task performance data to track whether subjects were on-task or distracted. From the data we only selected the blank phases (see Figure 1) as input for the ANN, because blank phases were assumed to involve most distracted thinking. As only half of the trials were followed by a thought-probe we solely considered these trials for analysis and classification.

Table 1. Correlations between the features. Bold values represent significant correlations ($p < 0.05$).

	PS	RE	FD	BR	BD	PP
PS	1	**-0.08**	**0.03**	**-0.09**	**-0.13**	**-0.03**
RE	**-0.08**	1	**0.06**	**0.05**	0.03	**0.10**
FD	**0.03**	**-0.24**	1	**-0.12**	**-0.08**	0.01
BR	**-0.09**	**0.05**	**-0.12**	1	**0.60**	**-0.05**
BD	**-0.13**	0.03	**-0.08**	**0.60**	1	**-0.04**
PP	**-0.03**	**0.10**	0.01	**-0.05**	**-0.04**	1

Extracting features. As input for the ANN, we extracted six features from the collected data: *mean pupil size* (PS), *rehearsal effort* (RE), *fixation duration* (FD), *blink duration* (BD), *blink rate* (BR), and *partial performance* (PP) on a single trial (i.e., percent correct). All features were z-score standardized ($\mu = 0$, $\sigma = 1$) by-subject prior to training the network. We decided to exclude blink rate from the classification analysis because it was correlated with blink duration ($r = 0.60$; see Table 1). We favored blink duration over blink rate because its values were more normally distributed. For details on how we computed the features we refer to the sections below. An overview of the feature extraction procedure can be found in Figure 1.

Mean pupil size. We first made sure that each blank phase was equally long by removing all pupil size values after 2 seconds from the start of the blank phase. The mean pupil size was subsequently computed for each blank phase by averaging the pupil size time series in each blank phase. The pupil size time series used for this calculation were corrected for gaze position [see 11].

Rehearsal effort. We determined rehearsal effort for each blank phase by counting the number of correct fixations in each blank. Correct fixations were defined as fixations on locations where previously in the trial a storage target was shown. To account for within-trial differences, we divided the number of correct fixations by the total amount of fixations in each blank phase. The resulting score was z-score standardized by-subject.

Fixation duration & Blink duration. The duration of fixations and blinks were determined by calculating the difference between the offset and onset of fixations and blinks in a blank phase. The software of the eye-tracker provided the time stamps for the onset and offset. When the end of a fixation or blink fell outside of the blank phase period, we did not cut off the duration but still regarded the full fixation or blink duration.

Blink rate. We computed the blink rate by counting the number of blink onsets in each blank phase.

Partial performance. The partial performance was calculated by counting the number of correct responses in each trial and dividing that by the span of the trial.

Algorithm. In this study, we created two 'types' of ANNs: full and lesioned. The full network received input from all features and therefore included five input neurons $\{x_i \mid x_1, \ldots, x_5\}$. Lesioned networks excluded one of the features and were designed to determine the contribution of a feature to classification by comparing a lesioned network to the full network. We created in total five lesioned networks, each excluding one of the features, and each network contained four input neurons. All networks had a single hidden layer with four neurons and one output neuron. We did try networks with two hidden layers and with different amounts of neurons. However, performance on the more complex ANNs did not result in significant improvement of classification performance. Therefore, we chose to the simpler network with one hidden layer over the two-layered alternatives.

All hidden and output neurons in full and lesioned networks used a *logistic (sigmoidal) activation* function. The logistic function takes input 'x' and transforms it into values between 0 and 1. We chose to use logistic function as it has been shown to perform well in the context of binary classification [see 12].

$$f(x) = \frac{1}{1 - e^{-x}} \tag{1}$$

As error function we applied *binary cross-entropy*,

$$E(X, Y) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log o(x_i) + (1 - y_i)\log(1 - o(x_i)) \tag{2}$$

where 'n' is the total amount of training examples, $Y = \{y_1, \ldots, y_n\}$ the true target labels (0 or 1), $X = \{x_1, \ldots, x_n\}$ the input, and $o(x)$ the output of the network (between 0 and 1). When the output of the network is close to the real target label, the resulting error from the function is close to zero. The rationale of choosing cross-entropy as error function is that it take the binary characteristics of the data into account and therefore provides a more optimal solution compared to other error functions such as mean squared error [see e.g., 13].

To optimize the weights in our networks we used *mini-batch stochastic gradient descent* with *Nesterov accelerated gradient* [15] as backpropagation algorithm,

$$v_t = \gamma v_{t-1} - \eta \frac{1}{B}\sum_{b=0}^{B-1} \frac{\partial E(W_{t-1}, \mathbf{m_b})}{\partial W_{t-1}} \tag{3}$$

$$W_t = W_{t-1} - \gamma v_{t-1} + (1 + \gamma) v_t \tag{4}$$

where weights 'W' are updated for every mini-batch $\mathbf{m_b}$ of 'B' training examples with a learning rate η. To improve the stability and convergence of the gradient descent, we used NAG to update the weights with respect to the direction of the previous weight updates. This was done by adding a momentum factor γv_{t-1} to the previous weight W_{t-1}, where momentum γ is a constant and v_{t-1} is the velocity of the gradient in the previous iteration *t*. To understand NAG, one can interpret the gradient as a ball that runs down a hill (i.e., the error/loss curve). When the ball moves in the right direction, it gains momentum. If there is a 'bump' in the curve (i.e., local minimum), the momentum will make sure that the ball can continue to progress. Because the movement of the gradient is calculated by predicting its path, the ball can also slow down when the slope of the curve goes up (i.e., when error is expected to increase). The result is that the ball will roll to the bottom of the curve, where the error is at its minimum (i.e., global minimum), without getting stuck in local minima and without overshooting the global minimum by rolling up the curve again.

Training and testing. Training and test sets were derived by performing *stratified 10-fold cross validation* on the full data set containing a total of 5374 examples. The resulting folds each had a training set with 4835 examples and a test set with 539.

Every network (i.e., full and lesioned) for each fold was trained independently for 500 epochs. The weights (W) were initialized at a random value between -0.1 and 0.1 and the bias was set to 1. We trained every network with a mini-batch size (B) of five and used a fixed learning rate (η) of 0.0001 and a momentum (γ) of 0.9.

Target labels for on-task thought examples were coded as 1, distracted thought examples were coded as 0. Because the network outputs probability values, we interpreted output values smaller than 0.51 as a prediction for class 0, and values equal or greater than 0.51 as a prediction for class 1. We chose 0.51 as threshold, because 51% of the examples in our data set were on-task.

3 Results

3.1 Logistic regression – how predictive are our features?

Before turning to the ANNs, we first wanted to check how predictive the individual features are of on-task vs. distracted thought. We fitted our features (excluding blink rate) as fixed effects on a logistic regression model with our subjects as random intercepts. Test results of this model are shown in Table 2.

We found that trial-to-trial partial performance (PP), mean pupil size (PS), rehearsal effort (RE), and blink duration (BD) were significant predictors of being on-task or distracted in the blank phase (all $p < 0.05$). The length of fixations (FD) was not significant, suggesting that fixation duration is indiscriminant of on-task and distracted states. While an increase in performance, pupil size, and rehearsal effort was found to be predictive of on-task thought, longer blinks were associated with being distracted. Partial performance was the strongest predictor, with one standard deviation (*SD*) increase in performance making it 1.77 times more likely to be on-task. One *SD* in-

crease in mean pupil size made it 1.18 times more likely to be on-task, and 1.09 times for rehearsal effort. To put these numbers in perspective, if the chance of being on-task is 50 percent, performing one *SD* better makes this chance 50 * 1.77 = 88.5 percent. For pupil size, rehearsal effort, and blink duration, this is 59, 54.5, and 46.5 percent respectively. Therefore we argue that partial performance is the most practically significant predictor. Although much less, eye-tracking features are also predictive for on-task vs. distracted thought.

3.2 ANN results – how well can we predict on-task and distracted states?

Now we know how predictive our features are, we can look at how the ANNs performed. All the important statistics are described in Table 3.

We found that all ANN classifiers, except for the lesioned network excluding partial performance, were able to predict on-task vs. distracted thought above chance at around 60%. Noticeably, the differences in classification performance between the full network and lesion networks excluding eye-tracking features were only very small. As the logistic regression model already showed that the influence of the eye-tracking features was small, this suggests that the contribution of eye-tracking features to the ANN classifier was minimal.

We observed a drop in classification accuracy when partial performance was excluded (i.e., L. PP classifier). The accuracy of the L. PP classifier was just above chance at 52.87%, and had an area under the receiver operating characteristic curve (AUC) of 0.539. The AUC is a measure of diagnostic ability and takes the ratio between true and false positives into account. An AUC of 0.539 means that amount of true and false positives was very similar, suggesting that the classifier did not learn the problem and potentially performed random guessing. More detail on the behavior of the classifier can be found in Figure 2 (right). Here, we see that the classifier made similar predictions on both on-task and distraction examples. Most of the output values were centered at 0.48, which happens to be the probability of encountering an distracted thought example (i.e., 0.487). Therefore, the classifier did not learn to classify on-task and distraction based on the data. Instead, it learned the underlying probability of distracted thought examples.

Table 2. Results from the logistic regression model with subjects as random intercept. Features are interpreted as significant when $p < 0.05$. Values in the e^{β} column are interpreted as "*x* times more likely with one unit increase in *feature.*"

Features	β	*SE*	e^{β}	*z*	*p*
PS	0.169	0.035	1.18	4.840	<0.001
RE	0.086	0.031	1.09	2.759	0.01
FD	0.028	0.031	1.03	0.887	0.37
BD	-0.069	0.031	0.93	-2.198	0.03
PP	0.573	0.035	1.77	16.441	<0.001

Table 3. Classification performance, collapsed over folds, of full and lesioned ANNs after training for 500 epochs. We reported the overall accuracy (acc, in %), area under the ROC curve (AUC), sensitivity to predicting distracted thought (DT), and specificity (i.e., accuracy of predicting on-task thought (OTT) on where DT is the true label). Full = all features. L. = lesioned.

	Overall acc (%)	*AUC*	*OTT acc (%)*	*DT acc (%)*
Full	60.31	0.623	53.26	67.76
L. PS	60.87	0.621	51.81	70.44
L. RE	60.61	0.622	52.35	69.33
L. BD	60.70	0.623	53.22	68.61
L. FD	60.48	0.623	52.79	68.61
L. PP	52.87	0.539	77.30	27.03

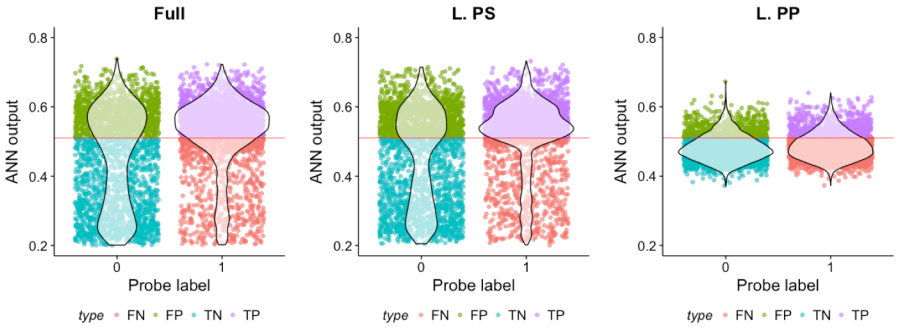

Fig. 2. The figures above show all output values and its density (i.e., shaded area) of the full ANN (left), the pupil size lesioned ANN (L. PS), and the trial performance lesioned ANN (L. PP). The value of the prediction is displayed on the y-axis, and the true probe labels are displayed on the x-axis. The red horizontal line shows the threshold used to classify a predicted value as on-task (value >= 0.51) or distracted (value < 0.51). The colors of the predictions show if the prediction is a true positive (TP; purple), true negative (TN; blue), false positive (FP, green), or a false negative (FN; red).

When we look in more detail at the classification performance of the full and eye-tracking lesioned classifiers, we found that the accuracy on on-task thought examples was much better compared to distracted thought examples. Accuracy on on-task examples was on average 69.10%, while for distracted thought examples it was only just above chance at 52.85% (see Table 3).

Now knowing that classification performance on distraction was poor, we wanted to examine why this was the case. We turned to the raw thought-probe data, and assessed the accuracy of the classifiers on each distraction response category (see Table 4). We observed that performance of the full and eye-tracking lesioned classifiers was below chance on examples where subjects engaged in mental elaboration (47.95% on

average) and in task-related, but interfering thought (44.31% on average). On the other hand, performance on the remaining distraction categories (i.e., external distraction, mind-wandering, and inattentiveness) was relatively good with accuracies up to 70%. It is likely that the classifier confused mental elaboration and task-related interference examples for on-task examples. This makes sense, because the thought content in these distraction categories is more closely related to the task in comparison to the other categories.

Table 4. Overview of the classification accuracy on different thought-probe categories (in columns) for all ANN classifiers (in rows). The bottom row 'n' gives the total amount of examples for each category in the full data set.

	On-task	Mental elaboration	Task-related interference	External distraction	Mind-wandering	Inattentiveness
Full	67.76	48.43	45.04	62.72	65.15	69.63
L. PS	70.44	47.19	42.83	61.82	63.00	69.63
L. RE	69.33	47.52	44.10	62.42	63.81	68.69
L. BD	68.61	48.51	44.88	64.42	63.81	69.16
L. FD	68.61	48.10	44.72	61.81	64.61	68.69
L. PP	27.03	74.63	74.02	79.10	86.33	83.64
n	2612	1210	635	330	373	214

3.3 ANN results – individual differences

The last thing we examined was the performance of the full ANN classifier on the different subjects in this study. We found that accuracy was above chance (accuracy > 0.51) for most of the subjects (n = 26, total = 32, 81.3%), and relatively good (accuracy > 0.7) for 21.8% of the subjects (n = 7). Notwithstanding, we found accuracies below chance of a handful, but substantial number of subjects (n = 6, 18.8%).

As performance of the classifier differed between subjects, we explored these differences in more detail. First, we examined whether the ratio of on-task thought/distracted thought reports of each subject correlated with accuracy. Because the classifier performed better on on-task examples, it might be that differences in the amount of on-task reports impacted the performance of the classifier. We found a small correlation ($r = 0.21$), but this was not significant ($p = 0.25$). Therefore, there was likely no relationship between the ratio of on-task/distraction reports and the performance of the classifier.

Second and last, we investigated if individual differences in the features were related to changes in classification accuracy. We calculated the mean value of each feature (excluding blink rate) for on-task and distracted thought examples for each subject. Subsequently, we calculated a difference score by subtracting the mean feature value of distracted thought examples from the on-task examples. Correlating these difference scores with classification accuracy revealed that differences in partial performance were strongly correlated with classification accuracy ($r = 0.79$, $p < 0.001$). In addition, we found a moderate positive correlation for rehearsal effort ($r =$

0.35, $p = 0.045$). Therefore, it is likely that the individual differences in classification performance were caused by variation in partial performance between on-task and distracted thought examples, and to a smaller extent by differences in rehearsal effort. Surprisingly, we found no evidence of a relationship between classification accuracy and across subject differences in pupil size ($r = 0.03$, $p = 0.86$). This was unexpected, because previous research [8] and our logistic regression model showed that pupil size was a reliable predictor. The ANN classifier was therefore inconsistent with this work.

4 Discussion

The aim of this study was to classify on-task and distracted thought examples with eye-tracking related and performance related features from a demanding task. The results showed that we were able to classify on-task and distracted though above chance (60% accuracy) on the basis of input from both eye-tracking (i.e., mean pupil size, rehearsal, blink duration, and fixation duration) and performance features (i.e., trial-to-trial partial performance). Noticeably, classification performance dropped to chance level when trial-to-trial partial performance was excluded from training and testing. This suggests that performance of our full network, receiving input from all features, was largely carried by partial performance. We conclude that the eye-tracking features only contributed little to the performance of the classifiers.

On-task thought was easier to classify than distracted thought. Detailed investigation of the classifiers' predictions on individual distracted thought categories showed that performance had stark differences. Whereas performance on mind-wandering, external distraction, and inalertness examples was relatively good (i.e., accuracy > 0.6), performance on mental elaboration and task-related interference was poor (i.e., accuracy < 0.5). It is likely that the latter distracted thought categories were confused for on-task examples. Taking into account that they were also the most prevalent distracted thought categories, it is implied that they contributed to the low performance on distracted thought examples.

Exploring the performance of the full classifier on the individual subjects showed that accuracy was above chance for the majority of subjects (i.e., 81%), and above 70% accuracy for a substantial group of subjects (i.e., 25%). We found that individual differences in the decrement (i.e., on-task – distracted thought) in partial performance and rehearsal effort were correlated with the differences in classification accuracy. Small decrements in both features were associated with chance level or below chance level performance. Importantly, we found no such relationship between the ratio of on-task and distracted thought examples in individual subjects and the classification performance. This means that the accuracy of the classifier on individual subjects can only be explained by differences in the feature data. From these results, we conclude that partial performance, and rehearsal effort to a lesser extent, contributed to successful classification.

Comparing the results of this study to other work is difficult, since there are only few mind-wandering classification studies, and these studies have used different feature sets and methodologies. Nevertheless, comparisons can be made on parts of the

results. We found that our classifier underperformed in comparison to other work. Other classification studies in the literature have reported average or median classification accuracies of 72 to 81% [see 6,8,15]. It should be noted that these studies did not include behavioral features such as trial-to-trial performance, but instead used only eye-tracking features [8] or fMRI-based features in combination with eye-tracking [6]. Both studies found that eye-tracking features contributed significantly to classification performance. Although it is difficult to determine why the present study did not identify such a contribution, it is possible that the challenging and dynamic nature of our task caused more noise and variance in the eye-tracking measures, and therefore made it more difficult to learn a good decision boundary.

Another possible explanation for the relatively weak contribution of eye-tracking features is that the content of distracted thought in our study was different from the other mentioned classification studies. Most subjects reported experiencing distracting thoughts as a result of the stimuli from the processing task (i.e., mental elaboration) in our study. In previous work using the same data set [9], we showed that this category of thought was not associated with a different mean pupil size compared to on-task. Similar results were found for task-related interference, while typical mind-wandering categories (i.e., mind-wandering and inattentiveness) were associated with a significantly lower mean pupil size [4]. Mind-wandering categories, however, only constituted 11% of all reports. Tasks with lower cognitive demand have found to involve more mind-wandering [e.g., 16], and given that the other mentioned classification studies used simpler tasks, this may at least explain why mean pupil size was a better predictor in their studies.

While different distracted thought categories affected mean pupil size differently, it may highlight the strength of eye-movement features. The advantage of eye-movement, such as fixations, saccades, and blinks, is that they are ubiquitous in most tasks [15]. For instance, the fixations on and saccades towards locations on the display where targets were presented (i.e., rehearsal effort) are cues that subjects rehearsed the targets in our paradigm. The absence of such eye-movement behavior means that the subject is not performing the task, and is therefore off-task/distracted. Although rehearsal effort was not significantly different from on-task for all distracted thought categories, there was a similar trend towards lower rehearsal effort visible for each distracted thought category. This highlights the potential for eye-movement features as predictors for distracted thought in general, and may also explain why rehearsal effort showed up in the individual differences analysis.

5 Conclusion and Future Directions

At the beginning of this paper we raised the research question: How are eye-tracking features and task performance related to distracted thought? From this study we conclude that trial-to-trial performance is the strongest predictor of distracted thought. Although eye-tracking related features have shown great potential, this study seems to indicate that the nature of the task and the content of distracted thought can greatly affect this potential. Rehearsal effort based on eye-movement behavior was found to

be the most promising eye-tracking feature. Although speculative, we suggest that eye-movement features are independent of the content of distracted thought and may therefore provide a more generic feature for classifying distracted thought. An interesting avenue for future research is to explore the eye-movement data in more detail, to determine which eye-movement features are relevant for classification. For example, recent research has shown that beta-process hidden markov models are able to extract dynamic patterns from time series examples [18] Applying such methods to eye-movement data may provide more insight in which patterns in the time series determine on- or distracted thought behavior.

Acknowledgements

This research was supported by a grant from the European Research Council (MULTITASK – 283597) awarded to Niels Taatgen. We would like to thank Marco Wiering for his comments and suggestions in the early stages of this research.

References

[1] J. Smallwood and J. W. Schooler, "The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness," *Annu. Rev. Psychol.*, vol. 66, no. 1, pp. 487–518, 2015.

[2] J. C. McVay and M. J. Kane, "Drifting from slow to 'D'oh!': working memory capacity and mind wandering predict extreme reaction times and executive control errors.," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 38, no. 3, pp. 525–549, 2012.

[3] Y. Weinstein, "Mind-wandering, how do I measure thee with probes? Let me count the ways," *Behav. Res. Methods*, no. December, 2017.

[4] N. Unsworth and M. K. Robison, "Pupillary correlates of lapses of sustained attention," *Cogn. Affect. Behav. Neurosci.*, vol. 16, no. 4, pp. 601–615, 2016.

[5] J. Smallwood, K. S. Brown, C. Tipper, B. Giesbrecht, M. S. Franklin, M. D. Mrazek, J. M. Carlson, and J. W. Schooler, "Pupillometric evidence for the decoupling of attention from perceptual input during offline thought," *PLoS One*, vol. 6, no. 3, 2011.

[6] M. Mittner, W. Boekel, a. M. Tucker, B. M. Turner, A. Heathcote, and B. U. Forstmann, "When the Brain Takes a Break: A Model-Based Analysis of Mind Wandering," *J. Neurosci.*, vol. 34, no. July, pp. 16286–16295, 2014.

[7] T. Foulsham, J. Farley, and A. Kingstone, "Mind wandering in sentence reading: decoupling the link between mind and eye.," *Can. J. Exp. Psychol.*, vol. 67, no. 1, pp. 51–9, 2013.

[8] R. Grandchamp, C. Braboszcz, and A. Delorme, "Oculometric variations during mind wandering," *Front. Psychol.*, vol. 5, no. FEB, 2014.

[9] S. Huijser, M. K. van Vugt, and N. A. Taatgen, *The Wandering Self: Tracking the Mind Wandering State in a Complex Working Memory Task*. Manuscript submitted for publication, 2017.

[10] D. Stawarczyk, S. Majerus, M. Maj, M. Van der Linden, and A. D'Argembeau, "Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method," *Acta Psychol. (Amst).*, vol. 136, no. 3, pp. 370–381, 2011.

[11] J. Brisson, M. Mainville, D. Mailloux, C. Beaulieu, J. Serres, and S. Sirois, "Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers.," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1322–1331, 2013.

[12] M. Minsky and S. Papert, *Perceptrons*. Oxford, England: M.I.T. Press, 1969.

[13] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, 2002.

[14] L. M. Silva, J. Marques de Sá, and L. A. Alexandre, "Data classification with multilayer perceptrons using a generalized error function," *Neural Networks*, vol. 21, no. 9, pp. 1302–1310, 2008.

[15] Y. Nesterov, "A method for solving a convex programming problem with convergence rate O(1/k2)," *Sov. Math. Dokl.*, vol. 27, no. 2, pp. 372–376, 1983.

[16] R. Bixler and S. D'Mello, "Toward Fully Automated Person-Independent Detection of Mind-Wandering," in *International Conference on User Modeling, Adaptation, and Personalization*, 2014, pp. 37–48.

[17] J. Rummel and C. D. Boywitt, "Controlling the stream of thought: Working memory capacity predicts adjustment of mind-wandering to situational demands," *Psychon. Bull. Rev.*, vol. 21, no. 5, pp. 1309–1315, 2014.

[18] E. B. Fox, M. C. Hughes, E. B. Sudderth, and M. I. Jordan, "Joint modeling of multiple time series via the beta process with application to motion capture segmentation," *Ann. Appl. Stat.*, vol. 8, no. 3, pp. 1281–1313, 2014.