Evaluation of a Personalized Method for Proactive Mind Wandering Reduction

Robert Bixler[1], Kristopher Kopp[2], and Sidney D'Mello[1,2]

[1]Deparment of Computer Science and [2]Department of Psychology, University of Notre Dame
{rbixler, kkopp, sdmello}@nd.edu

Abstract. We report on a project with the goal of creating a proactive system that attempts to reduce the propensity to mind wander (MW) by optimizing learning conditions (e.g., text difficulty and value) for individual learners. Our previous work had shown that supervised classification based on individual attributes could be used to detect the learning condition with the lowest MW rates. Here we test the model by comparing MW rates for the predicted optimal conditions to MW rates from a random control condition or in the condition with the overall best MW rate across all learners. Our results suggest that our method is better than these non-adaptive alternatives in certain contexts.

Keywords: engagement, mind wandering, affect, machine learning

1 Introduction

Learner models are at the core of intelligent tutoring systems (ITS). The development of ITSs has been influenced by cognitive learner models [1,2], and in recent years there has been a rise in ITSs that have been informed by affective models [3,4,5]. The cognitive-affective state of engagement is of particular interest for this project. Engagement has been defined as an enjoyable state of involvement in a learning activity or task with focused attention and intense concentration [3]. Engagement is necessary for learning since learners have to attend to information in order to learn. *Mind wandering* (MW) pertains to instances where engagement is disrupted and learners involuntarily shift their attention from their task towards unrelated thoughts. MW can be detrimental to learning [6, 7] because of this lapse in attention. Thus, to facilitate learning, it is important to develop systems that can either sustain engagement by reducing the propensity of MW behaviors or respond when a learner becomes disengaged due to MW. Not all learners exhibit the same MW behaviors when placed in the same learning environments [8]. Some learners experience lower MW rates compared to others depending on the context of the learning activity. For example, in a situation where the learning materials are considered difficult some individuals may be able to sustain attention and remain engaged, while others may disengage as their attention drifts towards thoughts unrelated to the task. With this in mind, we have begun developing a method that adapts the learning environment according to measures of individual attributes in an effort to reduce MW behaviors during a learning session. Our intention is to select optimal learning materials based on these

measures, with the purpose of reducing the propensity to MW. For example, learners would be assessed for attributes such as reading comprehension or scholastic aptitude and would then be placed in a learning environment and provided with materials based on those attributes with the goal of reducing the propensity to MW. The goal of this paper is to evaluate the performance of such a system by comparing our method to two non-adaptive alternatives.

1.1 Related Work

A variety of learner models have been employed in ITSs since their inception. Examples of cognitive models include: knowledge tracing models [9, 10], item response theory models [11], and knowledge space models [12]. Recent research of alternatives to cognitive models includes affective models [13, 14], meta-cognitive models [15], and models of disengagement [16] (see [17] for a review of recent models). Advancing the groundwork laid by studies that have investigated the relationship between affect and learning [3], [see 18 for a review, 19], recent research along these lines has led to the development of **Reactive** affect-sensitive ITSs that attempt to sense affective states related to learning and respond accordingly [20, 21, 22]. One example of this type of system is Affective AutoTutor [23]. This system detects the cognitive-affective states of the learner (i.e., boredom, confusion) based on conversational modeling, facial cues, and body language and alters the dynamics of the tutoring session through dialog moves designed to address specific affective states.

Although there are no analogous reactive systems that respond to MW, there have been attempts to develop automatic MW detectors. Drummond and Litman [24] used acoustic-prosodic features extracted from learners' utterances during a spoken learning task to discriminate episodes of low "zoning out" from episodes of high "zoning out", obtaining an accuracy of 64%. With a similar goal in mind, Bixler and D'Mello [25] recently attempted to automatically detect MW during reading on a computer screen using eye movements. They were able to detect MW with an accuracy of 72% (expected = 61%). A similar system, called GazeTutor [4], used an eye tracker to detect when users looked away from the screen for an extended period of time, which was taken to imply attentional disengagement. Although GazeTutor didn't definitively detect instances of MW, it attempted to re-engage learners with interventions when attentional disengagement was detected. Thus, research is steadily moving towards systems that are able to identify and respond appropriately to MW with the goal of sustaining engagement and improve learning.

Conversely, **Proactive** strategies attempt to facilitate affective states that would be beneficial for learning or avoid states that would be detrimental for learning. One example of a system that used a proactive strategy is ConfusionTutor [26], which attempted to induce a state of confusion during learning as there has been evidence that suggests a positive correlation between learning gains and confusion [27].

1.2 The Current Project

We recently took a step towards developing a proactive strategy to reduce MW by selecting learning materials that lead to reduced MW rates for individual learners [8].

MW rates were estimated with learner responses to auditory probes while learners read instructional texts on a computer screen. Each text was either an easy or difficult version and was manipulated to have either low or high value with respect to its weight on a subsequent test. Each learner read a total of four texts: one of each combination of difficulty and value. Supervised learning methods were used to build models that used individual attributes to predict the texts that would result in the lowest MW rate for that learner. Each model was built on data from the other learners (i.e., N – 1) and was then applied to the learner that was held out. The best models were moderately successful, resulting in an accuracy of 64% (expected = 53%). The next step, and the focus of our current research, is to further investigate how effective our method is at personalizing the learning environment in order to reduce MW.

There are many ways to evaluate the effectiveness of a personalized system. Several empirical evaluation methods are mentioned by Chin [28], such as experimental comparisons between systems with and without learner models or evaluating the accuracy of each learner model. Gena [29] covers strategies for evaluating user-adaptive systems, which includes additional strategies such as user-centered evaluation through questionnaires and interviews, observational evaluation through user observation and log files, and predictive evaluations such as expert reviews. Similar evaluation methods are suggested specifically for ITSs by Mark and Greer [30]. Due to the early nature of this project, we opted for a preliminary analysis that takes advantage of existing data in lieu of a more time consuming experimental study.

The present work describes an investigation of the effectiveness of our method to prevent MW [8]. We used existing data which identified the MW rate of each learner for four different learning materials that varied in difficulty and value. To evaluate our method, we then selected a MW rate for each learner based on the model's prediction of the learning materials with the lowest MW rate (i.e., individual best). We then compared these to MW rates derived from two non-adaptive alternative methods. The first alternative was to determine the learning materials with the lowest MW rate on average across all learners and select those learning materials for each learner (i.e., overall best). The second alternative was to simply select learning materials for each learner at random (i.e., random).

2 Data and Methods

What follows is a description of the data collection and analyses for the current project. For a more detailed description of data collection, see [8].

2.1 Data Collection

Undergraduate students (N = 187) from two U.S. universities learned about research methods topics from four texts (i.e., experimenter bias, replication, causality, and dependent variables) presented on a computer screen. Each text contained 1500 words on average (SD = 10) and were split into 30-36 pages. The difficulty and value of each text was manipulated. The difficulty manipulation consisted of presenting either

an easy or a difficult version of each text. Value was manipulated based on the weight assigned to each text on a subsequent posttest. Learners read all four texts with one text in each one of the four conditions: 2 (difficulty: easy vs. difficult) × 2 (value: high vs. low). The success of the manipulations was confirmed with self-reports of the perceived difficulty and perceived value of the texts (see [31]). During the task, learners' MW was measured along with several individual attributes.

Mind Wandering was measured through auditory probes (i.e. a beep) on nine pseudorandomly chosen "probe pages" per text, a standard and validated method for collecting online MW reports [6]. The MW rate for each text was then obtained by computing the proportion of "Yes" responses to probes.

Individual Attribute measures were collected for use as features in our models. The following measures were collected: (a) reading comprehension, (b) reading fluency, (c) working memory ability, (d) interest in research methods, (e) general boredom proneness, (f, g) boredom in academic situations (underwhelmed and overwhelmed), (h) scholastic aptitude, and (i) prior knowledge. Scores of all measures were standardized by school to alleviate any large discrepancies due to demographic differences between schools.

Procedure. Learners began the task by proceeding through one of two 24 item multiple choice pretests (counterbalanced between pre and posttest across all learners) and several individual attribute measurements. After being given instructions on the learning task, they studied four texts (one at a time) on a page-by-page basis, using the space bar to navigate forward. The title of the text and the corresponding weight of the test questions (value manipulation) were explicitly presented before each text. After learners studied all four texts, they were presented with the remaining 24 item posttest and remaining individual attribute measures.

2.2 Supervised Machine Learning

We used measurements of the individual attributes to predict the learning materials (in terms of difficulty and value) that would result in the least amount of mind wandering using supervised learning. Models were built for 34 machine learning algorithms from the WEKA machine learning software [32]. These included lazy-learners, Bayesian models, decision trees, support vector machines, regression models, etc. There were two additional parameters. The first parameter was the minimum allowable difference (i.e., threshold) between a learner's standardized MW rate for the best and worst materials (i.e., a difference of .0, .25, or .5 standard deviations between the highest and lowest MW rates). The second parameter was the specific classification task. The task was to classify the optimal learning materials between low and high difficulty texts, low and high value conditions, or any of the 4 conditions. Leave-one-person-out cross validation was used to evaluate each data set. Models were built on all learners except for a hold out learner and then tested on the hold out learner; this process was repeated for all learners. This method ensures that the training and testing set for each model are learner-independent. The Kappa statistic was taken as the measure of classifier accuracy. A kappa value of 1 indicates perfect agreement, while a kappa value of 0 indicates agreement was no better than chance.

2.3 Comparison Analysis

The best performing models for each classification task were identified based on the highest kappa. The best model for both the difficulty and difficulty/value classification tasks was built with a decision stump classifier, while the best model for value was built with a simple logistic classifier. These models were then used to assess how our method of assigning materials to learners would perform compared to non-adaptive methods. To illustrate how each MW rate was computed for the comparison, consider a hypothetical situation with 4 learners being compared in the difficulty classification task (Table 1). Individual best MW rates are based on model predictions; in this example, the model predicted that the best materials would be the difficult texts for learners 2 and 3, and the easy texts for learners 1 and 4 (note that the model erred for learners 1 and 2). Overall best MW rates are the MW rates for each learner with the materials that resulted in the lowest MW rate on average across participants; these are the easy texts for this example. Random MW rates are the MW rates for each learner with materials chosen at random; in this example, learner 2 is randomly assigned difficult texts, while learners 1, 3, and 4 are randomly assigned easy texts. Note that in this case, both the overall best and individual best conditions predicted the materials with the lowest MW rate for half the learners, which resulted in comparable average MW rates of about 0.45.

Table 1. MW rates (proportions of yes to total probe responses) for 4 hypothetical learners by classification (easy and difficult) and comparison groups.

Learner	Easy	Difficult	Individual Best	Overall Best	Random
1	0.61	0.39	0.61 – Easy	0.61 – Easy	0.61 – Easy
2	0.28	0.56	0.56 – Difficult	0.28 – Easy	0.56 – Difficult
3	0.67	0.44	0.44 – Difficult	0.67 – Easy	0.67 – Easy
4	0.22	0.61	0.22 – Easy	0.22 – Easy	0.22 – Easy
Average	**0.44**	**0.50**	**0.46**	**0.45**	**0.52**

3 Results

Table 2 lists the average standardized MW rates for each of these conditions based on the complete data set. Our initial step was to assess the accuracies of the classification results when considering all four types of learning materials: difficulty (easy and difficult) × value (low and high). We compared the MW rates of the best performing model (i.e., at the threshold of .25 *sd*'s) which resulted in a kappa of .11 (observed accuracy of 34%, expected accuracy of 26%). The MW rates were significantly lower for the individual best condition compared to the random condition, $t(140) = -2.1$, $p = .04$, but not significantly different from the overall best condition.

We then collapsed across value and then difficulty and conducted similar analyses for each. Value, at the threshold of .25 *sd*'s, resulted in a kappa of .16 (observed accuracy of 59%, expected accuracy of 51%). The MW rates in the individual best condition were not significantly different from either the random or the overall best condi-

tion. Difficulty at the threshold of .5 *sd*'s, resulted in a kappa of .24 (observed accuracy of 64%, expected accuracy of 53%). The MW rates in the individual best condition were significantly different from the random condition, $t(97) = -2.4$, $p = .02$, but not different from the overall best condition .

Table 2. Standardized MW rate means by classification task (standard deviations in parentheses). Lower numbers are preferred.

Classification Task	Individual Best	Overall Best	Random	*N*
Difficulty × Value	-.01 (.87)	-.03 (.87)	.13 (.93)	141
Value	-.05 (.79)	-.01 (.80)	-.01 (.81)	141
Difficulty	.07 (.72)	.09 (.75)	.17 (.75)	98

These preliminary results show that the models built on a small suite of individual attributes chose learning materials for each learner that were optimal in terms of resulting in the least amount of MW when compared to placing learners into a random learning condition except when collapsing across value. However, we were unable to choose materials with reported instances of MW that were statistically less than those chosen in the overall best condition across all learners.

We next wanted to take a close look at those individuals whose best model condition was different than the overall best condition to gain further insight into how the mind wandering behaviors differ between these conditions (see Table 3). The analyses described above were repeated after removing learners with the same individual best and overall best condition. For example, if the model predicted a learner should be given low difficulty materials, which is the overall best condition, then that leaner would not be included in the following analysis. For each analysis, the sample size was considerably culled resulting in low power, however, the results of significance are still reported. When considering all four types of learning materials (i.e., difficulty × value value) at the threshold of .25 *sd's*, the MW rates for the individual best condition were higher than the rates for the overall best condition, $t(40) = .799$, $p = .43$. When considering only value at the threshold of .25 *sd's*, the rates for the individual best condition were lower than the overall best condition, $t(51) = -1.5$, $p = .13$. When considering only difficulty at the threshold of .5 *sd's*, the rates for the individual best condition were lower than the overall best condition, $t(18) = -91$, $p = .38$.

Table 3. Standardized MW rate means by classification task for learners that differed on MW rates for the individual best and overall best conditions (standard deviations in parentheses)

Classification Task	Individual Best	Overall Best	Random	*N*
Difficulty × Value	.20 (.78)	.09 (.82)	.28 (.92)	41
Value	.07 (.81)	.17 (.80)	.12 (.82)	52
Difficulty	.14 (.84)	.24 (.97)	.17 (.90)	19

This second analysis shows that when the individual best condition differs from the overall best condition of all learners, there are some drastic differences in the amount of MW rates. When collapsing on value or difficulty (separately), the individual best condition outperforms the overall best condition. However, when considering the difficulty × value classifications, this trend is reversed where the overall best produces the least amounts of mind wandering.

4 Discussion

The goal of this project is to take strides towards creating a personalized learning environment in which a learner is provided with materials that reduce the propensity to MW. While there have been a few encouraging projects that attempt to take such proactive steps toward enhancing the learning experience by adapting the learning environment [see 33 for a review], this project's focus on attempting to proactively sustain engagement by reducing the likelihood that learners would MW based on a rather small number of individual attribute measures is novel. We showed that our method performs either better than or at least as well as two non-adaptive alternatives for choosing learning materials that will lead to a reduced MW rate. This is an initial step towards developing a system sensitive to learners' needs in terms of sustaining engagement. The next step would be to implement an experiment to test the generalizability of the claim that the method described here is, in fact, an effective method to incorporate into a preventative learning environment. Another possibility is to assess an expanded set of individual attribute measures. An exploration of additional measures could determine a specific set of features that are best able to predict a condition with an optimal MW rate.

Two limitations are apparent. First, it is possible that learners reported MW rates incorrectly, which could decrease the accuracy of our method. However, learner self-reports are used extensively in previous studies as there is not currently a good alternative for tracking MW [6]. Second, these findings are based on learners reading texts on research methods in a laboratory setting. Future work could boost claims of generalizability by incorporating different topics and other modes of information delivery.

This research takes a step towards tailoring a learning environment in order to reduce the rate of MW and potentially increase engagement. Systems exist that are sensitive to various states of the learner and take a reactive approach by adapting to the needs of the learner in a variety of contexts [21, 22, 23]. This project takes a proactive approach to addressing the needs of the learner by assessing their attributes and identifying learning materials that would potentially produce the least amount of MW. This method need not be limited to addressing MW behaviors during a leaning session. It would be beneficial for future work to assess how this method could be applied to addressing other cognitive affective states, such as boredom or confusion, which also have an influence on learning.

Acknowledgment. This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958). Any opinions, findings and

conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

References

1. Graesser, A. C., Olney, A., Haynes, B. C., & Chipman, P. (2005). AutoTutor: a cognitive system that simulates a tutor that facilitates learning through mixed-initiative dialogue. In C. Forsythe, M. L. Bernard, & T. E. Goldsmith (Eds.), Cognitive systems: Human cognitive models in systems design. Mahwah, NJ: Erlbaum.
2. VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading?. *Cognitive Science*, *31*(1), 3-62.
3. Baker, R., D'Mello, S., Rodrigo, M., & Graesser, A. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies, 68 (4), 223-241.
4. D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies, 70*(5), 377-398.
5. Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: Recognizing and responding to student affect. *International Journal of Learning Technology, 4*(3/4), 129-163.
6. Mooneyham, B. W., & Schooler, J. W. (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *67*(1), 11.
7. Szpunar, K. K., Moulton, S. T., & Schacter, D. L. (2013). Mind Wandering and education: from the classroom to online learning. *Frontiers in psychology*, *4*.
8. Kopp, K., Bixler, R., D'Mello, S. K. (In Press). Identifying Learning Conditions that Minimize Mind Wandering by Modeling Individual Attributes. Lecture Notes in Computer Science 8474, Springer-Verlag, pp. 94-103
9. Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education (IJAIED)*, *8*, 30-43.
10. Mitrovic, A. (2012). Fifteen years of constraint-based tutors: what we have achieved and where we are going. *User Modeling and User-Adapted Interaction*, *22*(1-2), 39-72
11. Millán, E., & Pérez-De-La-Cruz, J. L. (2002). A Bayesian diagnostic algorithm for student modeling and its evaluation. User Modeling and User-Adapted Interaction, 12(2-3), 281-330.
12. Falmagne, J. C., Cosyn, E., Doignon, J. P., & Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. In *Formal concept analysis* (pp. 61-79). Springer Berlin Heidelberg.
13. Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, *19*(3), 267-303.
14. D'Mello, S., & Graesser, A. (2012). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(4), 23.
15. Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. International Journal of Artificial Intelligence in Education, 16(2), 101-128.

16. Baker, R. S., Mitrović, A., & Mathews, M. (2010). Detecting gaming the system in constraint-based tutors. In User Modeling, Adaptation, and Personalization (pp. 267-278). Springer Berlin Heidelberg.
17. Desmarais, M. C., & d Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction, 22(1-2), 9-38.
18. D'Mello, S. K. (2013). A Selective Meta-analysis on the Relative Incidence of Discrete Affective States during Learning with Technology, Journal of Educational Psychology.
19. Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, *29*(3), 241-250.
20. Calvo, R. A., & D'Mello, S. K. (2012). Frontiers of affect-aware learning technologies, IEEE Intelligent Systems, *27(6)*, 86-89
21. Baker, R. S., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., ... & Rossi, L. (2012, June). Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. In *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 126-133).
22. Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, *19*(3), 267-303.
23. D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., ... & Graesser, A. (2008, June). AutoTutor detects and responds to learners affective and cognitive states. In *Workshop on Emotional and Cognitive Issues at the International Conference on Intelligent Tutoring Systems*.
24. Drummond, J. & Litman, D. (2010). "In the Zone: Towards Detecting Student Zoning Out Using Supervised Machine Learning," Intelligent Tutoring Systems, Part Ii, Lecture Notes in Computer Science 6095, V. Aleven, et al., eds., Springer-Verlag, pp. 306-308
25. Bixler, R., & D'Mello, S. K. (in press). Toward Fully Automated Person-Independent Detection of Mind Wandering. In User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg.
26. D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153-170.
27. D'Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition & Emotion*, *25*(7), 1299-1308.
28. Chin, D. N. (2001). Empirical evaluation of user models and user-adapted systems. *User modeling and user-adapted interaction*, *11*(1-2), 181-194.
29. Gena, C. (2005). Methods and techniques for the evaluation of user-adaptive systems. The knowledge engineering review, 20(01), 1-37.
30. Mark, M. A., & Greer, J. E. (1993). Evaluation methodologies for intelligent tutoring systems. *Journal of Artificial Intelligence in Education*, *4*, 129-129.
31. Mills, C., & D'Mello, S. K. (in prep). How Do Extrinsic Value and Difficulty Impact Engagement: An Experimental Approach.
32. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*. 11, 1, 10–18.
33. D'Mello, S. K. & Graesser, A. C. (in press). Feeling, Thinking, and Computing with Affect-Aware Learning Technologies. In Calvo, R. A., D'Mello, S. K., Gratch, J., & Kappas, A. (Eds.) Handbook of Affective Computing. Oxford University Press.