Partner Keystrokes can Predict Attentional States during Chat-based Conversations

Vishal Kuvar
University of Minnesota
kuvar001@umn.edu

Lauren Flynn
University of Minnesota
flynn598@umn.edu

Laura Allen
University of Minnesota
lallen@umn.edu

Caitlin Mills
University of Minnesota
cmills@umn.edu

ABSTRACT

Computer-mediated social learning contexts have become increasingly popular over the last few years; yet existing models of students' cognitive-affective states have been slower to adopt dyadic interaction data for predictions. Here, we explore the possibility of capitalizing on the inherently social component of collaborative learning by using keystroke log data to make predictions across conversational partners (i.e., using person A's data to make prediction about if person B is mind wandering). Log files from 33 dyads (total N = 66) were used to examine: a) how mind wandering (defined here as task-unrelated thought) during computer-mediated conversations is related to critical outcomes of the conversation (trust, likability, agreement); b) if task-unrelated thought can be predicted by the keystrokes of one's partner; and c) how much data is needed to make predictions by testing various window-sizes of data preceding task-unrelated thought reports. Results indicated a negative relationship between task-unrelated thought and perceptions of the conversation, suggesting that attention is an important factor during computer mediated chat conversations. Finally, in line with our hypothesis, results from mixed effects models showed that one's level of task-unrelated thought was predicted by the keystroke patterns of their conversational partner, but only using small window sizes (5s worth of data).

Keywords

mind wandering, chat, keystrokes, task unrelated thought

1. INTRODUCTION

Imagine you are messaging with a classmate about a homework assignment that is due in your programming class later that day. You exchange rapid messages back and forth, discussing how to debug the problem. You send a last message, but your partner does not immediately reply. Until this moment, your attention had been almost entirely focused on the conversation. But now, in this moment of silence, your attention is captured by thoughts of going to the grocery store once you're finished. You think about how crowded it will probably be, then brainstorm what you want to cook later, and start to think about how you wish you had a sandwich right now. At some point a few minutes later, your friend messages you back and you suddenly realize how far your mind had wandered away from the conversation you two were having.

This example illustrates a critical feature of our attention – namely, that it is constrained by the actions of the people we interact with. In the context of conversation, for example, we are influenced by the content of the messages that our partner sends but also by a variety of more subtle behaviors, such as the timing of the responses themselves. Such timing information is commonly captured via log files in educational technologies, and there is a long history of using this information to predict cognitive and affective states during learning [8]. However, these approaches typically rely on log file information for a particular student to make a prediction about that same student's cognitive state. As our example above illustrates, it may be the case that the behaviors of a conversational partner can provide important information about students' cognitive states that would not otherwise be apparent. With only access to your log data, we would not know why you stopped messaging your partner – was it because you were bored, gave up, or got distracted? Knowing your partner's behaviors helps answer this question perhaps even better than your own.

Here, we expand traditional modeling approaches in the EDM community by examining the predictive power of partner log data to predict attentional states. We designed a computer-mediated conversation task and logged keystroke data from pairs of students while they talked. Periodically, the students were asked to provide self-reports of their attentional states, operationalized here as task-unrelated thought (TUT). Rather than using each student's keystrokes to predict their own attentional state, we test whether they can be predicted from the keystrokes of their partners. Assessing the feasibility of using partner data to predict cognitive states is particularly timely today where interactions amongst individuals are increasingly occurring online and may continue in this direction with the advent of large language model based chatbots (e.g., Chat-GPT). It is therefore important that we consider new methodologies that rely on

V. Kuvar, L. Flynn, L. Allen, and C. Mills. Partner keystrokes can predict attentional states during chat-based conversations. In M. Feng, T. Käser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 217–223, Bengaluru, India, July 2023. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.8115697

numerous sources of log data beyond those of the individual student, which can provide opportunities to model and respond to student attention.

1.1 Related Work

1.1.1 Task-Unrelated Thought

TUT tends to occur around 20-30% during computerized reading [9], 30-40% during online lectures [22], and 20% while interacting with an intelligent tutoring system [16]. Importantly, TUT frequency has consistently demonstrated a negative relationship with affective valence [18] and learning outcomes [9, 25]. Given the frequently negative consequences of TUT on learning, researchers and educational technology developers have placed a strong emphasis on the development of models that can detect when a student has gone off task based on log data that can be readily integrated within a system. These models have relied on a variety of different sources of data to date, such as reading times, eye gaze, and EEG signals [11, 10]. These detectors can then be used to increase adaptivity and personalization in educational technologies. For example, recent work has shown that a TUT-sensitive intervention was effective for promoting long-term comprehension compared to a control group who did not receive the interventions at the moment they needed them [17].

Despite the substantial body of work on TUT during learning, it has rarely been examined in collaborative contexts, such as computer-mediated communication (CMC) or interactive learning contexts where chats are the most common form of communication among students (or between teachers/bots and students). One exception is recent work demonstrating that TUT occurs quite frequently when students are chatting with one another on computers in separate locations; instances of TUT were also correlated negative affective valence and other variables during the chat, providing initial evidence that it might be an important indicator of chat outcomes [4]. However, this study currently exists in isolation, leaving large gaps in our understanding of if and how TUT matters during conversations.

Given the nascent work in this area, our first research questions center around if and how such instances of TUT relate to perceived conversational outcomes; that is, what is the benefit of knowing whether students are off-task, and is it predictive of outcomes we care about in collaborative learning? Answering these questions will provide a baseline for future work in the context of EDM – namely motivating why we should consider modeling attention in the context of student computer mediated chats. A few variables that are of particular interest along these lines are likability and trust [19, 23]. However, trust is often difficult to measure directly or in real-time, so proxy stealth measurements that are linked to trust could provide "early warnings" for interventions. Indeed, for any chat-based system to be effective, these variables will be critically important to understand and detect early on so that students don't disengage before it's too late.

At the same time, if TUT is predictive of key outcomes, then we also need to understand effective ways to model it as chat conversations unfold. In our context, we are focused on understanding which behaviors, that can be readily extracted in chat data, may be used to predict cognitive states – particularly ones that capture the inherent social interdependence that exists within dyadic chats. This may be quite timely to explore given that CMC – especially remote, real-time chats – is becoming increasingly used in educational settings.

1.1.2 Keystroke Data

We chose to use keystroke data to explore this question given past work showing that keystroke log files are able to provide fine-grained temporal information about students' language production. For instance, the number of keys a student presses at the beginning of a writing task can provide insights into the degree to which their ideas were developed before they began the task. Similarly, a high number of backspaces may indicate that the student was revising their ideas in the moment. Keystroke features such as these have been linked to numerous factors related to learning, such as emotions [1, 3], reflective evaluation [24], and the quality of written product itself [14, 1, 5].

Predictive models using keystroke data have predominantly focused on writing tasks completed by single students, such as argumentative essays (see [6] for a review). However, there is some work in the CMC literature that examines the role of message timing in conversational success; for example, research indicates that shorter pauses with fewer keystrokes are associated with increased trust in your conversational partner [13]. These prior studies provide a foundation for work using keystrokes in CMC contexts; however, many questions remain unanswered. Relevant to the current work, how might the keystrokes of our partners relate to our own attentional states? As illustrated above, it is likely that the rhythm of our conversational partner may have a direct influence on our own attentional states; however, research is still needed to provide a more formal account of how partners' behaviors relate to cognitive and affective states.

1.2 Overview and Novelty of Current Work

There is no shortage of educational technologies that can detect and respond to an individual's cognitive states. Still, relatively few studies have leveraged the inherent interdependence between individuals in social contexts to inform such technologies. As collaborative learning becomes increasingly popular, understanding these links may open new doors to predictive modeling in collaborative tasks. Towards this goal, we expand the traditional application of log files to make cross-partner predictions of attention in a dyadic conversation from readily available keystroke log files from 33 dyads. In doing so, we answer the following research questions: a) how TUT during computer-mediated conversations is related to critical outcomes of the conversation (trust, likability, agreement); b) if TUT can be predicted by the keystrokes of one's partner; and c) how much data is needed to make predictions by testing various window-sizes of data preceding task-unrelated thought reports.

2. METHODS

2.1 Data Collection

2.1.1 Participants

We collected data from participants using an online platform called Prolific, where participants were paid to engage

Table 1: Keystroke features and descriptions.

Feature Type	Keystroke Feature	Description	Mean (Std.)	
			5s	15s
Non-message	Verbosity	Number of keystrokes in the window	20.34 (6.229)	47.98 (16.63)
	Backspace Frequency	Number of times the backspace key was hit in the window	0.002 (0.020)	0.005 (0.028)
	Maximum Latency	Maximum difference between two successive keystrokes in the window	5713 (4578)	15257 (10933)
	Median Latency	Median difference between two successive keystrokes in window	1152 (2616)	864.8 (2485)
Message	Inter-Word Time	Mean time between consecutive words in the recreated message	334.6 (120.9)	619.4 (272.3)
	Inter-Sentence Time	Mean time between consecutive sentence in the recreated message	354.3 (1194)	588.9 (1736)
	Number of Words	Number of words in recreated message (separated by space key)	2.563 (0.903)	5.997 (2.244)
	Maximum Sentence Length(# of Keystrokes)	Maximum number of words logged to type a sentence	18.79 (1.665)	41.33 (14.53)

in our chat-based study. All participants (n=218) locations were limited to the United States and the United Kingdom. Prolific has been shown to be a reliable source for data collection and can yield more diverse datasets in terms of participant background and age. Participants had a mean age of 34.016 (SD=11.45). 73.7% were female, 24.7% male, and 1.6% reported being non-binary. 79.1% participants were Caucasian, 11% Asian, 3.4% Hispanic, 1.1% Black, 0.5% American Indian and 4.9% as 'other.'

2.1.2 Chat Platform

We built our own online chat platform where two participants would be automatically matched and converse with each other while answering in-situ "thought probes" about whether they were experiencing TUT throughout the chat session. The chat was designed to be much like an large online discussion, where two students may be randomly paired with each other and asked to chat. The entire conversation lasted a total of 16 minutes. During the conversation, all keystrokes and their associated timestamps were logged. We attempted to keep the conversations relatively open-ended to mimic real-life chats between students. Each conversational pair was given one of three different instructions for the conversation to create diversity in the chats (i.e. so any findings could not be attributed to forcing a single type of chat/topic): 1) high constraints, where the participants were asked explicitly to learn about remedies for the common cold from one another; 2) low constraint, in which participants were asked to discuss medically relevant topics; 3) no constraints, where participants were simply asked to chat with each other. Note that this manipulation was not necessarily intended to lead to differences in our dependent variables (and we find no significant differences in the key variables across conditions); rather, we include it to test whether our findings generalize across multiple topics. However, for the sake of caution, we included topical condition as a "control" variable in all of the analyses presented.

After the task, participants were redirected to a survey page where they answered questions about their demographics, valence, and arousal. Valence measures how positive to negative participants feel, whereas arousal captures participant's level of activation, or how sleepy to active they feel. The survey also consisted of questions about trust, likability, and agreement the participant felt towards their partner.

2.1.3 Thought Probes

Six brief thought probes were administered pseudo-randomly throughout the duration of the conversation. Both participants saw the probes simultaneously about every two and a half minutes. The probe read: "On a scale of 7, please select a number that most reflects your attention on the current task right now. 1 being you are completely focused on task and 7 being you are not focused on the task at all." This thought probe method is the gold-standard in TUT research, particularly in educational contexts [9]. Although there are inevitably some limitations that come with using self-reports, this method has been validated numerous times in different contexts (lab settings, online research, classrooms). Results suggest that thought probes do not have a negative influence on task performance, and the responses have reliable correlates [21]. In our study, both conversational partners received the probe simultaneously. Message sending was disabled until the participant responded to the probe.

2.1.4 Trust, Likability, and Agreement Questionnaires

Participants answered questions about trust, likability, and agreement immediately after talking to their chat partner. An 11 item scale designed by McCallister [15] was used to measure trust. Likability was measured using a modified version of the Rysen Likability Scale (RLS; [20]). Out of the original 11 items on that scale, we chose five to include that were relevant to online interactions. Items on both questionnaires were reported using a 7-point Likert scale. Five questions were constructed to measure the agreement of chat perceptions between participants. An example of a question was "I was interested in my partner's viewpoint." Participants reported on a 9-point Likert Scale. Participants

answers were added for each scale and these sums were used in subsequent analyses.

2.2 Data Processing and Analyses

2.2.1 Data Cleaning

We collected 218 keystroke files. Due to a glitch in the server, 33 of the files logged keystrokes as "Unidentified." Additionally, four files contained fewer than 200 keystrokes. These 37 files could not be used for the feature extraction process and were dropped. Given that our second research question was based on interdependence, it was imperative that were able to align data from both conversational partners. However, given that we were unable to align for these same 37 participants, their respective conversational partners was also dropped. We also removed all pairs of participants who did not get a total of six probes due to an error in the triggering system (N = 39 pairs). The final total number of participants that could be used for analysis was 66 (33 pairs).

2.2.2 Window Creation

Our primary goal is to assess whether keystroke patterns can be used to predict the attention of someone's conversational partner. We thus needed to align keystroke patterns with thought reports in a time-sensitive manner. That is, we needed to extract keystroke data from a "window" leading up to the thought report (but not including the report itself). This windowed approach is commonly used for detecting TUT in real-time [11], but this is the first time it has been applied in a dyadic context, where we take data from one person to make a prediction about the other.

We created windows leading up to each thought probe using two window sizes that have been successful in prior research: 5s and 15s [11]. We created these windows in the time leading up to the probe, such that keystroke data extracted from the chat would predict their partner's future level of TUT. The window was defined by the nearest keystroke to the thought probe. That is, if a person did not type in the 5s window immediately preceding the thought probe, the algorithm would instead search for the closest keystroke and begin the window at this point. This approach was necessary given the dyadic nature of the task, where conversational partners often take consecutive turns. If the keystroke preceding the probe was typed outside of the window size, only that keystroke would be included in the window. This results in the features extracted from these windows to have low values, compared to when keystrokes are present.

2.2.3 Features

Once a window was defined, keystroke features were extracted and classified into two categories: message and non-message features (see Table 1). Message features require the recreation of the message within the window, whereas non-message features use the raw keystrokes. The non-message features that we selected were based on Bixler and D'Mello [3]. Messages in the window were recreated by processing the keystrokes in the window in a sequential manner. A space key indicated a new word. A period, exclamation mark, or question mark indicated the end of a sentence. If a backspace key was encountered in a window, the previously typed key would be deleted.

Table 2: Correlation matrix of self-reported TUT, arousal, and partner perception.

	Prop. TUT reports	Valence	Arousal	Trust	Likability
Valence	-.201				
Arousal	.117	.194			
Trust	-.372*	.083	.075		
Likability	-.261·	.311·	.122	.600**	
Agreement	-.312·	.193	.009	.742**	.521**

**p<0.01, *p<0.05, .p<0.1

2.2.4 Analytical Approach

The lme4 package [2] in R was used to create mixed-effects models. We extracted features from the raw keystrokes and used the standardized version of them as data. Models were fitted with random intercepts, with the participant acting as the random effect. This accounted for within-subjects variance in the responses. For this analysis, the independent variables were the individual keystroke features of a participant. The dependent variable was the response of their partner for the TUT probe. We report the unstandardized regression coefficient (B), p-value, 95% confidence intervals, and standardized β.

3. RESULTS

Given that TUT has not been studied often in the context of CMC, a major contribution of this work is evidence that participants' average level of TUT was 2.40. This indicates they were predominantly on task relative to the midpoint of the scale (3.5 on a 1 to 7 scale), but nevertheless went off task quite a bit (SD = 1.36). Participants also seemed to feel generally positive with an average affective valence of 3.53 (SD = .78), and they appear to moderately trust and agree with their conversational partners.

3.1 Does TUT relate to the perceptions of conversation?

TUT has a consistent negative relationship for affective valence and learning outcomes, but these are almost explicitly studied in individual tasks. Our first research question was thus to explore how levels of TUT relate to affective states as well as perceptions of the chat. These correlations help address a basic question: is TUT worth detecting in the context of conversations? Table 2 presents the Pearson correlation values between variables, where each person's own level of TUT was correlated with their self-reported affect and perceptions of the chat. Out of the 66 participants, only 62 were used to calculate these values. Data for the remaining four were missing. First, we replicated the typically observed negative relationship between TUT and affective valence positive [18]. Second, we also observed significant correlations between TUT and perceptions of the chat – namely, increased TUT was associated with less trust, likability, and agreement with your conversational partner.

3.2 Do keystrokes predict partner TUT and at what window size?

Table 3: Results of mixed effects models.

Predictors (Keystroke features)	Attention level of conversational partner									
	5s window					15s window				
	B	β	p	95% CI		B	β	p	95% CI	
				Lower	Upper				Lower	Upper
Intercept	3.02	0.00	<0.001	2.61	3.43	3.02	0.00	<0.001	2.61	3.43
Verbosity	0.17	0.08	0.04	0.00	0.35	0.08	0.04	0.42	-0.11	0.26
Backspace frequency	0.00	0.00	0.99	-0.16	0.16	-0.01	-0.01	0.81	-0.18	0.14
Maximum latency	0.08	0.04	0.31	-0.08	0.24	-0.02	-0.02	0.67	-0.20	0.13
Median latency	0.13	0.06	0.11	-0.03	0.29	0.07	0.07	0.05	-0.01	0.31
Inter-word time	0.10	0.05	0.20	-0.06	0.26	-0.03	-0.03	0.46	-0.23	0.10
Inter-sentence time	-0.17	-0.08	0.03	-0.33	-0.01	-0.06	-0.06	0.08	-0.30	0.02
Word count	0.15	0.07	0.08	-0.02	0.32	0.03	0.03	0.50	-0.12	0.24
Maximum sentence length (Keystrokes)	0.18	0.08	0.04	0.01	0.36	0.03	0.03	0.42	-0.11	0.25

Table 3 provides the full results for all regression models. For the 5s window, three keystroke features significantly predicted partner's level of TUT: verbosity, inter-sentence time and maximum sentence length. Verbosity was positively related to partner TUT, suggesting that when someone types for longer periods (with more keystrokes), their partner's mind is more likely to drift off-task. A similar relationship was observed between maximum sentence length and TUT. Taken together, these relationships indicate that when individuals produce longer messages, their partners may be more likely to go off-task, perhaps while waiting for their partner to complete their thought.

Unlike the keystroke features above, the inter-sentence time feature was negatively related to partner TUT, with a one standard deviation increase in inter-sentence length corresponding to a 0.17 decrease in TUT. The inter-sentence time feature provides information about when partners pause between the sentences they produce. Thus, it provides some context for the pauses in the chat rather than simply examining all pauses regardless of when they occur. The negative relationship between this feature and TUT suggests that certain types of pauses may be more or less important for a partner's TUT. In particular, if an individual pauses after drafting a full sentence, it is more likely the case that their partner now has a complete idea that they can reflect upon and respond to, rather than something more incomplete. This is a particularly compelling interpretation, given that overall pause times (latencies) were unrelated to partner TUT reports.

Importantly, all of the significant relationships that we observed were for keystroke features calculated at the 5s window, not at the 15s window. This suggests that the keystroke features were predictive of partner TUT rates, but only for those recorded immediately before the probe was delivered. We cannot make causal claims about why this is the case; however, a strong possibility is that the window sizes for keystroke features are sensitive to the specific type of conversation that is taking place. Here, students were asked to have a conversation while not engaging in any other tasks – this single-task paradigm resulted in relatively rapid turn-taking between the partners.

Finally, it is worth noting that even the significant models revealed a relatively small effect in terms of the variance explained by the fixed effects effects in our linear mixed effects regressions. The predictors verbosity, inter-sentence time, and maximum sentence length had a conditional R^2 value of 0.007, 0.006, and 0.007, respectively – leaving an opportunity to refine such models in future work.

4. DISCUSSION

Collaborative learning environments are inherently social. One person's actions will inevitably influence others. The current study leveraged this interdependence among individuals in a conversational setting in order to determine if log file data can be helpful for predicting cross-partner cognitive states. Our main hypothesis was that, in a conversational setting, one partner's keystrokes may be indicative of their partner's attentional state. Taken together, our results support this hypothesis – highlighting the idea that incorporating the interdependence between individuals into predictive models may be beneficial for adaptive educational technologies supporting collaborative learning. Specifically, we demonstrate that keystrokes are one such feature that can be leveraged to make these predictions in the context of computer-mediated conversations.

Verbosity, inter-sentence time, and maximum sentence length were the three keystroke features that were reliably predictive of partner TUT. However, it is worth noting that these features were only significant within relatively short (5s) window sizes. The window sizes at which keystroke features should be calculated are likely to depend on the context of the conversation taking place; thus, when examining keystroke data, researchers should extract features at multiple window sizes to determine which are most appropriate for their context. A second implication of our study for future research using keystroke data is the use of non-message and message features. We found that inter-sentence time was a reliable predictor of partner TUT ratings, but there

were no relations to overall pause time. This indicates that keystroke features may benefit from the addition of contextual information from the conversation itself. For example, pauses after highly emotional messages may reflect different processes than those after relatively mundane messages, such as making plans or asking simple questions.

Our study adds to the growing body of work suggesting that keystrokes are an indicator of cognitive states. Keystroke features, such as the ones extracted here, are readily available in most log files, yet are not commonly used in multimodal models. It may therefore be worthwhile adding this feature to increase predictive power in both individual and collaborative settings. Our paper is focused on the latter context; as such, we believe there may be particular promise in interactive group contexts as individual models may not always reveal the whole story: is someone interacting less because they are bored, frustrated, or confused? One's own data may not be the best way to answer this question; perhaps the person cannot get a word in because the conversation is moving too fast, or perhaps everyone else has stopped responding. Adding keystrokes to predictive models can thus allow for the social component to be included in a more explicit way, facilitating time-sensitive nudges or subtle feedback to conversational partners on their interactions.

A final set of findings emerged to suggest that TUT is worth monitoring in CMC. People seemed to experience TUT quite often during computer-mediated conversations, in line with previous work showing it is ubiquitous in almost all aspects of our lives, including educational activities [12, 25]. Not only does it happen often, these experiences do not appear to be particularly positive; increased levels of TUT were consistently and negatively related to trust, likability, and agreement amongst partners in ours study. This complements prior work that links TUT to negative affect and clinical conditions – underscoring a potential need to detect it and respond in educational technologies.

An interesting possibility to consider, particularly as chatbots (e.g., Chat-GPT) are likely to continue rapidly evolving, is how our findings can be expanded in a chatbot setting. With chatbots becoming more knowledgeable and accessible, a possibility that bots can be used in education cannot be ignored. Future work may consider exploring how chatbots mimicking keystroke patterns that are associated with lower levels of TUT may influence engagement, and thus learning outcomes [25]. There is already some evidence that predictive models of TUT (using one's own keystrokes) are accurate during dyadic CMC interactions, and that the results generalize to chatbot settings; expanding this work in more ecological and with multiple conversational participants' keystrokes would likely be fruitful (i.e. even beyond dyadic interactions to group interactions).

Like most research, ours is not without limitations. For example, we created our own chat platform and did not provide participants with an explicit learning goal. Although we took care to vary the topic, replicating our results under different goal conditions will be an important next step. We were also somewhat limited with sample size, limiting our scalability. Nevertheless, our analytical approach provides proof-of-concept for the usefulness of using partner data. Future research may also wish to improve our models by including content-dependent features, such as the conversational topic. These limitations and caveats notwithstanding, we believe that "attending to attention" [7] will be helpful in building technologies that can facilitate effective online collaboration.

5. REFERENCES

[1] L. K. Allen, M. E. Jacovina, M. Dascalu, R. D. Roscoe, K. M. Kent, A. D. Likens, and D. S. McNamara. {ENTER}ing the Time Series {SPACE}: Uncovering the Writing Process through Keystroke Analyses. Technical report, International Educational Data Mining Society, 2016. Publication Title: International Educational Data Mining Society ERIC Number: ED592674.

[2] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48, Oct. 2015.

[3] R. Bixler and S. D'Mello. Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13*, page 225, Santa Monica, California, USA, 2013. ACM Press.

[4] A. Colby, A. Wong, L. Allen, A. Kun, and C. Mills. Perceived group identity alters task-unrelated thought and attentional divergence during conversations. *Cognitive Science*, 47(1):e13236:1–23, 2023.

[5] R. Conijn, C. Cook, M. van Zaanen, and L. Van Waes. Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, 32(4):835–866, Dec. 2022.

[6] P. Deane and M. Zhang. Automated Writing Process Analysis. In D. Yan, A. A. Rupp, and P. W. Foltz, editors, *Handbook of Automated Scoring*, pages 347–364. Chapman and Hall/CRC, 1 edition, Feb. 2020.

[7] S. D'Mello, K. Kopp, R. E. Bixler, and N. Bosch. Attending to Attention: Detecting and Combating Mind Wandering during Computerized Reading. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*, pages 1661–1669, San Jose, California, USA, 2016. ACM Press.

[8] S. K. D'mello and J. Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.*, 47(3):1–36, feb 2015.

[9] S. K. D'Mello and C. S. Mills. Mind wandering during reading: An interdisciplinary and integrative review of psychological, computing, and intervention research and theory. *Language and Linguistics Compass*, 15(4):e12412:1–32, 2021.

[10] H. W. Dong, C. Mills, R. T. Knight, and J. W. Y. Kam. Detection of mind wandering using EEG: Within and across individuals. *PLOS ONE*, 16(5):1–18, May 2021. Publisher: Public Library of Science.

[11] M. Faber, R. Bixler, and S. K. D'Mello. An automated behavioral measure of mind wandering during computerized reading. *Behav Res*, 50(1):134–150, Feb. 2018.

[12] S. Hutt, K. Krasich, C. Mills, N. Bosch, S. White, J. R. Brockmole, and S. K. D'Mello. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Model User-Adap Inter*, 29(4):821–867, Sept. 2019.

[13] Y. M. Kalman, L. E. Scissors, A. J. Gill, and D. Gergle. Online chronemics convey social information. *Computers in Human Behavior*, 29(3):1260–1269, May 2013.

[14] A. D. Likens, A. Likens, L. K. Allen, and D. S. McNamara. Keystroke Dynamics Predict Essay Quality. In *Annual Meeting of the Cognitive Science Society*, page 6, July 2017.

[15] D. J. McAllister. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *The Academy of Management Journal*, 38(1):24–59, 1995.

[16] C. Mills, S. D'Mello, N. Bosch, and A. M. Olney. Mind Wandering During Learning with an Intelligent Tutoring System. In C. Conati, N. Heffernan, A. Mitrovic, and M. F. Verdejo, editors, *Artificial Intelligence in Education*, volume 9112, pages 267–276, Cham, 2015. Springer International Publishing.

[17] C. Mills, J. Gregg, R. Bixler, and S. K. D'Mello. Eye-Mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human–Computer Interaction*, 36(4):306–332, July 2021.

[18] C. Mills, A. R. Porter, J. R. Andrews-Hanna, K. Christoff, and A. Colby. How task-unrelated and freely moving thought relate to affect: Evidence for dissociable patterns in everyday life. *Emotion (Washington, D.C.)*, 21(5):1029–1040, Aug. 2021.

[19] E. Molleman, A. Nauta, and B. P. Buunk. Social Comparison-Based Thoughts in Groups: Their Associations With Interpersonal Trust and Learning Outcomes. *Journal of Applied Social Psychology*, 37(6):1163–1180, 2007.

[20] S. Reysen. CONSTRUCTION OF A NEW SCALE: THE REYSEN LIKABILITY SCALE. *Social Behavior and Personality: an international journal*, 33(2):201–208, Jan. 2005.

[21] A.-L. Schubert, G. T. Frischkorn, and J. Rummel. The validity of the online thought-probing procedure of mind wandering is not threatened by variations of probe rate and probe framing. *Psychological Research*, 84(7):1846–1856, Oct. 2020.

[22] K. K. Szpunar, N. Y. Khan, and D. L. Schacter. Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, 110(16):6313–6317, Apr. 2013.

[23] S. W. Uranowitz and K. O. Doyle. Being liked and teaching: The effects and bases of personal likability in college instruction. *Res High Educ*, 9(1):15–41, Mar. 1978.

[24] A. Wengelin, M. Torrance, K. Holmqvist, S. Simpson, D. Galbraith, V. Johansson, and R. Johansson. Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior Research Methods*, 41(2):337–351, May 2009.

[25] A. Y. Wong, S. L. Smith, C. A. McGrath, L. E. Flynn, and C. Mills. Task-unrelated thought during educational activities: A meta-analysis of its occurrence and relationship with learning. *Contemporary Educational Psychology*, 71:102098:1–18, Oct. 2022.