Generalizability of Face-Based Mind Wandering Detection Across Task Contexts

Angela Stewart
University of Notre Dame
384 Fitzpatrick Hall
Notre Dame, IN, 46556, USA
astewa12@nd.edu

Nigel Bosch
University of Illinois at Urbana-Champaign
1205 West Clark Street
Urbana, IL, 61801, USA
pnb@illinois.edu

Sidney K. D'Mello
University of Notre Dame
118 Haggar Hall
Notre Dame, IN, 46556
sdmello@nd.edu

ABSTRACT

We investigate generalizability of face-based detectors of mind wandering across task contexts. We leveraged data from two lab studies: one where 152 college students read a scientific text and another where 109 college students watched a narrative film. We automatically extracted facial expressions and body motion features, which were used to train supervised machine learning models on each dataset, as well as a concatenated dataset. We applied models from each task context (scientific text or narrative film) to the alternate context to study generalizability. We found that models trained on the narrative film dataset generalized to the scientific text dataset with no modifications, but the predicted mind wandering rate needed to be adjusted before models trained on the scientific text dataset would generalize to the narrative film dataset. Additionally, we analyzed generalizability of individual features and found that the lip tightener and jaw drop action units had the greatest potential to generalize across task contexts. We discuss findings and applications of our work to attention-aware learning technologies.

Keywords

Mind Wandering, Mental States, Attention Aware Interfaces, Cross-Corpus training.

1. INTRODUCTION

Consider a typical day when you were an undergraduate college student. Your first class is your favorite, so you are engaged in the lecture content and processing new information. In your next class, you watch a documentary about a subject that does not interest you, causing your attention to focus on unrelated thoughts of your social life, rather than processing the information in the video. Later, you work on a homework assignment that you find frustrating, leading to waning motivation. Towards the end of your day, you attend a chemistry lab, where you interact with a new educational game that teaches you the basics of chemical bonds. At some points you are enjoying the game, and thus engaged in deeply learning the content. However, you later become bored during a long period of repetitive gameplay, causing you to become distracted and miss important information. Throughout the day, your mental states (engagement, frustration, boredom) influenced your learning. Your learning experience could have been augmented with technology that responded to your changing mental state, thus assisting you in achieving the most effective learning experience.

Educational interfaces that detect and respond to student mental states are driven by work on cognitive and affective state modeling, which has been investigated for many years. For example, attention and affect has been modeled in educational tasks such as reading comprehension [6, 16, 28] and computerized tutoring [3, 19], among others. In general, there has been a plethora of work that has modeled a variety of mental states within specific educational tasks (e.g., [2, 15, 19]) to better understand these states and use that knowledge to facilitate student learning.

However, prior research has overwhelmingly investigated single task contexts, and has overlooked generalizability to different contexts. For example, models that track attention during reading might not generalize to lecture viewing, educational gaming, and so on. This makes it difficult to decouple task-specific effects from more fundamental patterns. In contrast, models that successfully generalize across multiple contexts should reveal observable signals (i.e. eye gaze, facial features, and physiology data) that are general, rather than task-specific. Models using such indicators will be key to developing adaptive technologies that are sensitive to student mental states and that can operate across a range of educational activities.

We report results on modeling mental states in a generalized way using mind wandering (MW) as a case study. MW is a ubiquitous phenomenon where thoughts shift from task-related processing to task-unrelated thoughts [15]. MW is estimated to occur anywhere from 20% - 50% of the time, depending on the person, task, and environmental context [23]. It is has also been associated with lower performance on a variety of educational tasks, such as reading comprehension [16] and retention of lecture content [29], thus impacting student learning.

As with work on other mental states, research on MW has largely failed to address models that generalize across contexts [6, 15]. MW detection has been investigated in reading comprehension [6, 16], narrative and instructional film comprehension [25, 26], and student interaction with an intelligent tutoring system (ITS) [19]. To our knowledge, no work has investigated MW detection with the goal of generalizability across task contexts.

We specifically investigate the generalizability of MW models across two task contexts - reading a scientific text and viewing a narrative film. These contexts were chosen because of their broad applicability to education in the classroom and online. For example, a documentary film could be shown in a sociology course or distance learning students could read instructional texts prior to engaging in an online discussion.

1.1 Related Work

Cross corpus training has been researched in a variety of classification problems, such as sentiment analysis [31] and acoustic-based emotion recognition [35]. Cross corpus training seeks to improve robustness of machine-learned models by leveraging multiple datasets in classifier training and testing. For example, Webb and Ferguson [32] applied cross corpus training techniques to characterize the function of segments of dialogue using automatically extracted lexical and syntactic features called cue phrases. Each extracted cue phrase was used to classify a segment of dialogue. They trained separate classifiers on two different datasets, and applied the classifier to the dataset on which it was not trained. They found the cross-training results were comparable to the results of training and testing on the same dataset (e.g. the best cross-trained classifier achieved and accuracy of 71%, compared to an accuracy of 81% when trained and tested on the same dataset). Additionally, they examined generalizability of the cue phrases across datasets by reducing the feature set to contain only cues present in both datasets. They found that reducing the feature set yielded slight improvements, and demonstrated the discriminative nature of a small number of features.

Zhang et. al. [35] similarly explored the use of multiple datasets for creating context-generalizable models. They built classifiers for valence and arousal on highly varied emotional speech datasets using a leave-one-corpora-out cross-validation technique. Additionally, they explored methods for data normalization (within each dataset and between datasets) and agglomeration of both labeled and unlabeled data. They found that, of their six emotional speech corpora, training on some subsets yielded higher accuracy than others. Their work suggested that careful selection of corpora best suited for training might yield better emotional speech recognition performance than an all-or-nothing approach to cross-corpus training.

Our work approaches cross-corpus modeling through detection of MW. A variety of studies have investigated MW detection during educational tasks, such a reading [15], interacting with an intelligent tutoring system (ITS) [19], or watching an educational video [26]. No work has focused on MW from a cross-corpus modeling perspective, to our knowledge, so we review the individual studies below.

Detection of MW from eye gaze features while reading has been amply investigated. For example, Bixler and D'Mello [4] built models to detect MW while students read texts about scientific research methods. This work made use of probe-caught reports (students respond yes or no to auditory thought probes of whether they were MW), instead of self-caught reports (students report whenever they catch themselves MW). Their analysis of eye gaze features showed that certain types of fixations were longer during MW. Specifically, they found that longer gaze fixations (consecutive fixations on a single word), first-pass fixations (fixations on a word during the first pass through a text), and single fixations (fixations on a word only fixated on once) were predictive of MW. In other work, Bixler and D'Mello [5] similarly used eye gaze features, but used self-caught reports of MW. They found that a greater number of fixations, longer saccade length, and line cross saccades were indicative of MW. Across studies on MW detection during reading, longer fixations were found to be indicative of MW [4, 15, 28], suggesting these features might generalize well.

Pham and Wang [26] similarly used consumer-grade equipment to detect MW while students watched videos from massively open online courses (MOOCs). They made use of heart rate, detected by monitoring fingertip blood flow, using the back camera of a smartphone (i.e., photoplethysmography). Their models achieved a 22% improvement over chance. Although their method for detecting MW could be implemented across a variety of tasks, the question of whether heart rate is indicative of MW across task contexts has not yet been investigated.

Hutt et. al. provided limited evidence of generalizability of MW detection across different learning tasks during student interaction with an ITS [19]. They employed a genetic algorithm to train a neural network using context-independent eye-gaze features and context-dependent interaction features (e.g., current progress within the ITS). They achieved an F_1 value of .490 (chance = .190). This work provided some evidence of generalizability because the visual stimuli and interaction patterns varied throughout. For example, students interacted with an animated pedagogical agent in a scaffolded dialogue phase and completed concept maps without the tutoring agent in another interaction phase. However, it is still unclear if their model would generalize to a broader range of tasks, particularly less interactive ones like reading or film viewing. Furthermore, their best-performing models used context-dependent features, which could prevent the detector from generalizing to a task where those features could not be used.

1.2 Novelty

Our contribution is novel in a variety of ways. First, we demonstrate the feasibility of building cross-context detectors of mental states, specifically MW. Further, previous work on MW detection has sometimes made use of context-specific features (e.g., reading times) that are not expected to generalize to other contexts [19, 25]. In contrast, our work detects MW using only facial features and upper body movement, recorded using commercial-off-the-shelf (COTS) webcams that are expected to generalize more broadly. Additionally, the use of COTS webcams support a broader implementation of MW detectors as webcams are ubiquitous in modern technology. This is in contrast to prior research that has used specialized equipment, like eye trackers [15, 19, 25] or physiology sensors [7], which students would likely not have access to.

2. DATASETS

This study makes use of narrative film [23] and scientific reading comprehension [22] datasets collected as part of a larger project. Here, we include details pertaining to video-based detection of MW.

2.1 Narrative Film Comprehension

Participants were 68 undergraduate students from a medium-sized private Midwestern university and 41 undergraduate students from a large public university in the Southern United States. Of the 109 students, 66% were female and their average age was 20.1 years. Students were compensated with course credit. Data from four students were discarded due to equipment failure.

Students viewed the narrative film *The Red Balloon* (1956), a 32.5-minute French-language film with English subtitles (Figure 1). The film has a musical score but only sparse dialogue. This short fantasy film depicts the story of a young Parisian boy who finds a red helium balloon and quickly discovers it has a mind of its own as it follows him wherever he goes. This film was selected because of the low likelihood that participants have previously seen it and because it has been used in other film comprehension studies [34].

measurement necessitate, but which in more enlightened countries are wholly unnecessary. This book is not prepared to meet the requirements and artificial restrictions of any syllabus, and it is not prepared to help students through any examination. I cannot help thinking, however, that if the type of student who puts more faith in learning formulae

Figure 1. A screenshot of the narrative film (left) and scientific text (right) are shown.

Students' faces and upper bodies were recorded with a low-cost ($30) consumer-grade webcam (Logitech C270).

Students were instructed to report MW throughout the film by pressing labeled keys on the keyboard. Specifically, students were asked to report a task-unrelated thought if they were "thinking about anything else besides the movie" and a task-related interference if they were "thinking about the task itself but not the actual content of the movie." A small beep sounded to register their report, but film play was not paused. After viewing the film, students took a short test about the content and completed additional measures not discussed further.

We recorded a total of 1,368 MW reports from the 105 participants with valid video recordings. In this work, we do not distinguish between the two types of MW, instead merging the task-unrelated thoughts and the task-related interferences, both of which represent thoughts independent of the content of the film.

2.2 Scientific Reading Comprehension

Participants were 104 undergraduate students from a medium-sized private Midwestern university and 48 undergraduate students from a large public university in the Southern United States. Of the 152 participants, 61% were female and their average age was 20.1 years. Participants were compensated with course credit. Data from eight participants were discarded due to equipment failure.

Students read an excerpt from *Soap-Bubbles and the Forces which Mould Them* [8]. Like *The Red Balloon* (Figure 1), we chose this text because its content would likely be unfamiliar to a majority of readers. The text contained around 6,500 words from the first chapter of the book. In all, 57 pages (screens of text) with an average of 115 words each were displayed on a computer screen in 36-pt Courier New typeface. The only modification to the text was the removal of images and references to them after verifying that these were not needed for comprehension.

Students who read the scientific text were instructed to report MW in the same way as those who watched the narrative film. They were instructed to report a task-unrelated thought if they were "thinking about anything else besides the task" and a task-related interference if they were "thinking about the task itself but not the actual content of the text." Participants completed a comprehension assessment after reading the text. We recorded a total of 3,168 MW reports from the 144 students with valid video recordings.

2.3 Self Reports of MW

MW was measured via self-reports in both studies, so it is prudent to discuss the validity of self-reports. We used self-reports because this is currently the most common approach to measure an inherently internal (but conscious) phenomenon [5, 15]. Self-reported MW has been linked to predictable patterns in physiology [30], pupillometry [17], eye-gaze [28] and task performance [27], providing evidence for the convergent and predictive validity for this approach. To improve the quality of self-reports, we encouraged students to report honestly and assured them that reporting MW would not in any way effect the credit they received for participation.

The alternative to using self-caught reports is using probe-caught reports, which require a student response to a thought-probe (e.g., a beep). We chose self-caught reports over the probe-caught because the probe-caught method can potentially interrupt the comprehension process (i.e., when participants report "no" to the probes). Interruptions are particularly problematic in the film comprehension task, as participants did not have control over the media presentation (i.e., no pausing or rewinding of the film). Furthermore, it is also unclear if a probe-caught report takes place at the beginning or end of MW, or somewhere in between. Conversely, self-caught reports are likely to occur at the end of a MW episode when the student became aware that they were not attending to the task at hand.

3. MACHINE LEARNING

We explored a variety of machine learning techniques for cross-context MW detection using the same approach to segmenting instances and constructing features for both datasets.

3.1 Segmenting Instances

Reports of MW were distributed throughout the course of the film viewing or text reading session. We created instances that corresponded to reports of MW by first adding a 4-second offset prior to the report. This was done to ensure that we captured participants' faces while MW vs. in the act of reporting MW itself (i.e., the preparation and execution of the key press). This 4-second offset was chosen based on four raters judgements of whether or not movement related to the key-press could be seen within offsets ranging from 0 to 6 seconds. Data was then extracted from the 20 seconds prior to the MW report. A window size of 20 seconds was chosen based on prior experimentation that sought to balance creating as many instances as possible (shorter window sizes) and having sufficient data in each window (longer window sizes) to detect MW.

We extracted "not MW" instances from windows of data between MW reports. The entire session (reading or video watching) was divided into 24-second segments (20 second windows of data and a 4 second offset as with the MW segments). Any segments

overlapping the 30 seconds prior to a MW report were discarded. We do not know precisely when MW starts, so we chose to discard instances overlapping the 30 seconds prior to MW reports, to separate students when they were actually MW from when they were not. We also discarded any segments overlapping a page turn (discussed in Section 3.2). All remaining segments were labeled Not MW. Our approach to segmenting instances is shown in Figure 2.

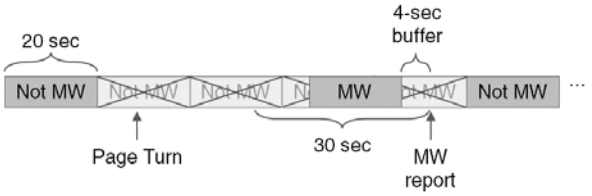

Figure 2. Illustration of the instance extraction method.

3.2 Instance Selection

A full accounting of the instance selection process is shown in Table 1. Our goal was to make the two data sets as similar as possible so that task-specific effects could be studied without additional confounds.

We first discarded any instances where there was less than one second of usable data in that time window. Data was not usable when the student's face was occluded due to extreme head pose or position, hand-to-face gestures, and rapid movements. Additionally, for the scientific reading dataset, we discarded instances that overlapped with page turn events. In prior experimentation, we trained a model to detect MW using only a binary feature of whether or not that instance overlapped a page turn boundary. MW was detected at rates above chance in this experimental model. Therefore, we concluded that including instances that overlapped page turn boundaries would inflate performance as the detector could simply be picking up on the act of pressing the key to advance to the next page.

After discarding instances using the method above, we matched the scientific reading and narrative film datasets on school (medium-sized Midwestern private university or large Southern public university), reported ethnicity, and reported gender. The scientific reading dataset was randomly downsampled to contain approximately the same number of students in each gender, race, or school category, as the film dataset. This participant-level matching on school, ethnicity, and gender was done to eliminate external sources of variance that could influence MW detection, potentially obfuscating task effects from population effects.

Finally, the datasets were downsampled to contain equal numbers of instances because the size of the training set is known to bias classifier performance [13]. We also downsampled the data to achieve a 25% MW rate in order to be consistent with research that suggests that MW occurs between 20% and 30% of the time during reading and film comprehension [6, 23]. Further, the MW rates of 30% and 14% obtained in these data are more artefacts of the instance segmentation approach rather than the objective rate, so resampling ensures a dataset that is more reflective of expected MW rates.

Table 1. An accounting of instance selection process

	Reading (% MW)	Film (% MW)
Base	7,267 (30%)	7,313 (14%)
Face Detected	7,266 (30%)	7,238 (14%)
Page Boundary	1,400 (36%)	N/A
Participant Matching	1,273 (35%)	N/A
Downsampling	1,100 (25%)	1,100 (25%)

3.3 Feature Extraction and Selection

We used commercial software, the Emotient SDK [36] to extract facial features. The Emotient SDK, a version of the CERT computer vision software [24] (Figure 3) provides likelihood estimates of the presence of 20 facial action units (AUs; specifically 1, 2, 4, 5, 6 ,7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, 26, 28, and 43 [14]) as well as head pose (orientation), face position (horizontal and vertical within the frame), and face size (a proxy for distance to camera). Additionally, we used a validated motion estimation algorithm to compute gross body movements [33]. Body movement was calculated by measuring the proportion of pixels in each video frame that differed by a threshold from a continuously updated estimate of the background image generated from the four previous frames.

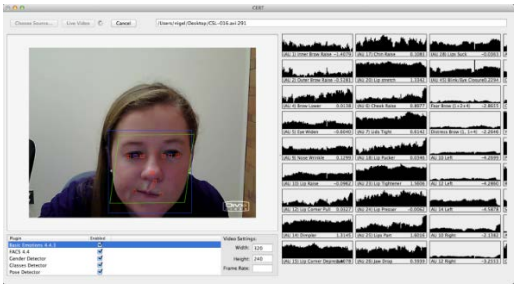

Figure 3. Interface demonstrating AU estimates detected from a face video.

Features were created by aggregating Emotient estimates in a window of time leading up to each MW or Not MW instance using minimum, maximum, median, mean, range, and standard deviation for aggregation. In all, there were 162 facial features (6 aggregation functions × [20 AUs + 3 head pose orientation axes + 2 face position coordinates + face size + Motion]). Outliers (values greater than three standard deviations from the mean) were replaced by the closest non-outlier value in a process called Winsorization [11].

We used tolerance analysis to eliminate features with high multicollinearity (variance inflation factor > 5) [1], after which, 37 features remained. This was followed by RELIEF-F [21] feature selection (on the training data only) to rank features. We retained a proportion of the highest ranked features for use in the models (proportions ranging from .05 to 1.0 were tested). Feature selection was performed using nested cross-validation on training data only. We ran 5 iterations of feature selection within each cross-validation fold (discussed below), using data from a randomly chosen 67% of students within the training set in each iteration.

3.4 Supervised Classification and Validation

Informed by preliminary experiments, we selected seven classifiers for more extensive tests (Naïve Bayes, Simple Logistic Regression, LogitBoost, Random Forest, C4.5, Stochastic Gradient Descent, and Classification via Regression) using the WEKA data mining

toolkit [18]. For each classifier, we applied SMOTE [9] to the training set only. SMOTE, a common machine learning technique for dealing with data imbalance, creates synthetic interpolated instances of the minority class to increase classification performance.

We evaluated the performance of our classifiers using leave-one-participant-out cross-validation. This process runs multiple iterations of each classifier in which, for each fold, the instances pertaining to a single participant are added to the test set and the training set is comprised of the instances for the other participants. Feature selection was performed on a subset of participants in the training set. The leave-one-out process was repeated for each participant, and the classifications of all folds were weighted equally to produce the overall result. This cross-validation approach ensured that in each fold, data from the same participant was in the training set or testing set but never both, thereby improving generalization to new participants.

Accuracy (recognition rate) is a common measure to evaluate performance in machine learning tasks. However, any classifier that defaults to predicting the majority class label of an imbalanced dataset can appear to have high accuracy despite incorrect predictions of all instances of the minority class label [20]. This is particularly detrimental in applications where detecting the minority class is of upmost importance. In our task, we prioritized the detection of MW despite the large imbalance in our dataset. Therefore, we considered the F_1 score for the MW label as our key measure of detection accuracy since F_1 attempts to strike a balance between precision and recall.

4. RESULTS

4.1 Cross-dataset Training and Testing

We trained three classifiers: one on the scientific text dataset, one on the narrative film dataset, and one on a concatenated dataset comprised of the first two. For each of the three training sets, the classifier that yielded the highest MW F_1 is shown in Table 2. We used leave-one-student-out cross validation for within-dataset evaluations. Conversely, to measure generalizability of the models across contexts we applied the classifier trained on scientific text data to the narrative film data, and vice versa. We compared our model to a chance model that classified a random 25% (MW prior proportion) of the instances as MW. This chance-level method yielded a precision and recall of .250 (equal to the MW base rate).

Table 2. Results for the models with highest MW F_1 for the within-data set validation (cross-training results in parentheses).

Training Set	Classifier	MW F_1	Precision	Recall
Scientific Text	Logitboost	.441 (.267)	.376 (.252)	.553 (.284)
Narrative Film	C4.5	.436 (.407)	.303 (.278)	.775 (.760)
Both	Logistic	.424	.314	.655

We calculated improvement over chance as (actual performance – chance)/(perfect performance – chance). All three models showed improvement over chance (25% for scientific text, 25% for narrative film, and 23% for the concatenated dataset) when trained and tested on the same dataset. When tested on the alternative dataset, the narrative film classifier generalized well to the scientific text dataset (21% improvement over chance). However, the scientific text model showed chance-level performance on the narrative film corpus (2% improvement over chance). The MW F_1 of the concatenated dataset model was simply an average of the MW F_1 score of the individual datasets when the instance predictions of the individual datasets are separated (.413 for the scientific reading dataset and .436 on the narrative film dataset). These results showed that the concatenated classifier does not skew towards predicting one dataset better than the other, but rather predicts both models with comparable accuracy.

Table 2 also shows precision and recall for each of the models. Across all models, recall was higher than precision, indicating a lot false positives. It is important to note the near chance-level recall and precision of the model trained on scientific reading data when applied to the narrative film data. The lack of improvement over chance for both recall and precision demonstrated the need to improve generalizability in both dimensions. Conversely, the cross-trained narrative film model had lower precision, but good recall, resulting in an improved MW F_1 score.

4.2 Classifier Generalizability

To address the negligible improvement over chance of the scientific text model when tested on the narrative film dataset, we repeated the training and testing using C4.5 as the classifier. The C4.5 classifier was chosen because it generalized better when trained on the narrative film dataset than the Logitboost classifier generalized when trained on the scientific text dataset. The results are shown in Table 3, where we note no notable improvement over the previous Logitboost classifier in Table 2 (change from .267 to .287 when tested on the narrative film dataset). Therefore, the lack of evidence for generalizability for the scientific text model could be due to overfitting to the training set, rather than classifier selection.

Table 3. Results (MW F_1) for the C4.5 classifier for within- and cross- validation.

Training Set	Within	Cross
Scientific Text	0.425	0.287
Narrative Film	0.436	0.407
Both	0.415	N/A

4.3 Prediction Threshold Adjustment

We further investigated the lack of generalizability of the scientific text model by considering the MW prediction rate. We compared the performance of both models on the narrative film dataset. Recall dropped considerably more than precision (Table 2; recall dropped from .775 to .284; precision decreased from .303 to .252). We hypothesized that recall decreased because of a difference in predicted MW rates (Table 4). In fact, the predicted MW rate in the narrative film data dropped from 64% to 28% when applying the scientific text model to the same data. This supported our hypothesis that the low recall was linked to lower predicted MW rates. Furthermore, 39% of the correctly classified instances (true positives and true negatives) were MW when applying the narrative film model to the narrative film data compared to 12% for the scientific text model applied to the same data. This demonstrated that the scientific text model was much more prone to missing MW instances, further supporting our hypothesis.

To address this, we adjusted the predicted MW rate of the scientific text model when applied to the narrative film dataset. The classifier outputs a likelihood of MW and we previously considered instances with likelihoods greater than .5 as MW. We adjusted that prediction threshold from .1 to 1 in increments of .1 (Figure 4) to investigate how changes in predicted MW rate (higher for lower thresholds) effected recall, and thus MW F_1.

Table 4. Predicted MW Rates.

Training Set	Within	Cross
Scientific Text	38%	**28%**
Narrative Film	**64%**	68%
Both	52%	N/A

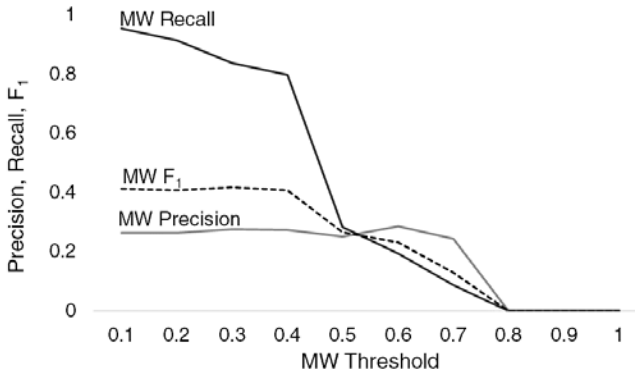

Figure 4. MW precision, recall, and F_1 as the prediction threshold varies for the scientific text model applied to the narrative film dataset.

We note that MW F_1 score degrades at a threshold of .5. We adjusted the threshold to .3 and yielded the results shown in Table 5. After adjusting the MW prediction threshold, both precision and recall of the narrative film data applied to the scientific text model showed comparable performance to the cross-trained narrative film model. It is important to note that the adjusted MW prediction threshold yielded a predicted MW rate of 76%, much higher than the MW rate of the dataset (25%). As with the generalized narrative film model, this reduced precision because the high predicted MW rate produced a large number of false positives.

Table 5. Results for models with highest MW F_1 (cross-training results in parentheses). Cross-training results for the scientific text model reflect a MW prediction threshold of .3.

Training Set	Classifier	MW F_1	Precision	Recall
Scientific Text	Logitboost	.441 (**.416**)	.376 (**.276**)	.553 (**.836**)
Narrative Film	C4.5	.436 (.407)	.303 (.278)	.775 (.760)
Both	Logistic	.424	.314	.655

4.4 Feature Analysis

We analyzed the facial features to further study generalizability by predicting MW with different subsets of the entire feature set. The C4.5 classifier was chosen for this feature analysis because of its consistency on both the scientific text model and concatenated dataset. Each subset consisted of the features (e.g., median, standard deviation) from one AU, or from face position, size, orientation, or motion. Since tolerance analysis was not used here, we only considered the minimum, maximum, median, and standard deviation aggregated features to prevent redundancy (e.g., between median and mean). For example, we used the minimum, maximum, median, and standard deviation feature values for AU5 (upper lid raiser) to predict MW. This approach was applied to the 20 AU subsets, as well as face position, size, orientation, and motion subsets. We generated the same cross-training configurations of in Section 4.1 (i.e., train on scientific text, test on narrative film, etc.). To rank the subsets of features on generalizability, we examined MW F_1 scores when testing on the alternative dataset only. For example, using the AU9 (nose wrinkle) subset, we investigated MW F_1 value of scientific text model applied to the narrative film dataset and the narrative film model applied to the scientific text dataset. Table 4 shows these results only for features that achieved a MW F_1 of greater than .250 (chance) on all dimensions (within dataset validation and cross-training). We selected features for further analysis if their MW F_1 was greater than .300 for both cross-training results. This value of .300 was used to filter out features that performed well on the within-dataset validation, but fell short on cross training. It also ensured that a feature performed better than chance on both cross-trained results (i.e., train on narrative film and test on scientific text, and vice versa), rather than only generalizing to one dataset. Using this criterion, only AU23 and AU26 showed notable improvement over chance.

We used the C4.5 classifier to generate the same models in Table 2 (train/test scientific text, train scientific text/test narrative film, etc.) using only the features from AU23 and AU26 (Table 7). None of these models (scientific text, narrative film, or concatenated) achieved a MW F_1 as high as those in Table 2, which used a combination of tolerance analysis and RELIEF-F to select features. This suggested that, while AU23 and AU26 might individually predict MW, when used together, their prediction power might be limited, compared to other feature selection techniques.

Table 6. MW F_1 score for within-data set validation with cross-data set scores (in parentheses).

	Training Set	
Facial Feature	**Scientific Text**	**Narrative Film**
AU4 (brow lowerer)	.378 (.278)	.398 (.395)
AU6 (cheek raiser)	.369 (.259)	.361 (.321)
AU9 (nose wrinkler)	.300 (.268)	.392 (.303)
AU14 (dimpler)	.303 (.267)	.383 (.376)
AU23 (lip tightener)	**.334 (.333)**	**.363 (.317)**
AU26 (jaw drop)	**.414 (.321)**	**.365 (.357)**
Face Height (size)	.322 (.256)	.339 (.289)
Face X (position)	.404 (.316)	.382 (.282)

Table 7. Results for models when only using the C4.5 classifier on AU23 and AU26.

Training Set	Classifier	MW F_1	Precision	Recall
Scientific Text	C4.5	.383 (.272)	.255 (.206)	.764 (.404)
Narrative Film	C4.5	.397 (.257)	.333 (.235)	.491 (.284)
Both	C4.5	.368	.271	.575

5. ANALYSIS

We developed automated detectors of MW using video-based features in the contexts of narrative film viewing and scientific reading. The generalizability of these models was dependent on corpora on which the model was trained and the rate at which the model predicts MW. In this section, we discuss our main findings and applications of this work. We also discuss limitations and future work.

5.1 Main Findings

We expanded on previous MW detection work through cross-context modeling. We trained three models on three datasets

(scientific text, narrative film, and a dataset concatenated from the two). We found each of these models (trained and tested on the same corpus) performed at a notable 23% to 25% improvement over chance. This demonstrated the feasibility of detecting MW on individual corpora. However, recall was greater than precision, indicating prediction of false positives. This should be considered when implementing MW detectors in educational environments where excessive prediction of student MW could be demotivating.

We investigated generalizability of the single-dataset models (i.e. scientific text or narrative film) by applying the model to the dataset on which it was not trained. The model trained on the narrative film dataset maintained performance when applied to the scientific text dataset (Table 2), providing some evidence for generalizability, but this performance was boosted by high recall (and comparatively low precision). Precision and recall (and thus MW F_1) were near chance-level when the model trained on the scientific text dataset was applied to the narrative film dataset, suggesting that the model might overfit to the scientific text training set.

We attempted to address this problem by applying the C4.5 classifier, as it comparatively generalized well when trained on the narrative film dataset. MW F_1 score for the scientific text classifier applied to the narrative film data again negligibly increased. This suggested that the training data (only scientific text) used was not appropriate for model generalization. This idea is supported by the performance of the narrative film model on the scientific text data (although detection of false positives is a limitation) and the notable improvement over chance (22% to 23%) for the concatenated dataset. The performance of both models suggested that there were discernable similarities between MW instances across the two datasets, which can be detected using our techniques.

In addition to training data, we also found that predicted MW rate effected model generalizability. We adjusted MW predictions according to a sliding threshold for the narrative film predictions obtained from the scientific text model. We found that relaxing the criteria for classifying an instance as MW (i.e. adjusting the likelihood prediction threshold from .5 to .3) yielded results comparable to the cross-trained narrative film model. However, this approach to increasing recall should be used with caution as it leads to increased likelihood of false positives. Perhaps in a real-time MW intervention scenario, a more balanced approach could be taken where the MW likelihood prediction is used to determine if a MW intervention is triggered (e.g., if the detector determines there is a 40% likelihood the student is MW, then there is a 40% chance a MW intervention is triggered).

We detected MW using individual feature subsets to ascertain whether certain face-based features (i.e. AUs, head orientation, position, size, and motion) generalize. We found two feature subsets (AU23 – lip tightener and AU26 – jaw drop) that showed a MW F_1 of at least .300 on both cross-trained models. It is notable that when looking at the generalizability of these features, they did not individually achieve MW F_1 scores as high as the best performing models in Table 2. This demonstrated the need for multiple features to work together to detect MW, rather than relying on a single feature. Furthermore, this showed that our method of feature selection (tolerance analysis and selecting a proportion of features using RELIEFF) was important to model performance.

5.2 Applications

The present findings are applicable to educational user interfaces that involve reading or film comprehension. Monitoring and responding to MW could greatly improve student performance on these tasks. Films and instructional texts play a major role in learning (both in the classroom and online). For example, films can give historical background on a time period being discussed in literature classes and instructional texts can supplement lecture content through textbooks or technical articles. Due to the relationship between MW and low task performance, user interfaces that detect and respond to MW in contexts where attention is key (i.e. education) would help students remain focused on their learning.

These findings are particularly promising for implementation in massively open online courses (MOOCs). Our method for detecting MW exclusively uses COTS webcams. These webcams are ubiquitous in today's computers and mobile devices; thus our work would integrate into a variety of learning environments without extra cost. Such a video-based detector of MW could feasibly respond to student MW through suggesting a student revisit text or video content, asking a reengaging question, or advising the student to take a break.

5.3 Limitations and Future Work

While we demonstrated techniques for modeling generalizability across task contexts, our work has a few limitations. First, precision is moderate, even on our best models. High predicted MW rates lead to high recall, but also more false positives. In this work, we chose to accept this tradeoff, with the goal of generalizability in mind. However, raising precision, while maintaining recall is key to task-generalizable MW detectors being successful in educational environments. Since MW is the minority class (25% of all instances), investigating skew-insensitive classifiers, such as Hellinger Distance Decision Trees [10], could improve precision.

Additionally, this work focuses exclusively on generalizability from the perspective of task context (viewing a narrative film vs. reading a scientific text). Claims of generalizability could be strengthened through MW detection across environments. Both the narrative film and scientific reading datasets were collected in a controlled lab setting. MW detection in the field, such as computer-enabled classrooms or the personal workstations of MOOC users, should be considered prior to implementation in such environments. Furthermore, student generalizability should be further examined. In this work, we detect MW in a student-independent way. However, participants were all of similar age and enrolled in college. Future work could examine the generalizability of our method for detecting MW in non-college-aged students, such as elementary students in a computer-enabled classroom or non-traditional students enrolled in distance learning courses.

5.4 Concluding Remarks

In this work, we showed evidence that generalizable detectors of MW can be created using video-based features. The corpora used to train models of MW and predicted MW rates both play a role in the model's ability to generalize and should be considered as work on cross-context MW generalization advances. This work advances the field of attention-aware interfaces [12] by demonstrating the feasibility of modeling MW across the educational contexts of reading a scientific text and viewing a narrative film. Our approach to detecting MW is the first step towards building interfaces that detect MW across multiple educational activities.

6. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF) (DRL 1235958 and IIS 1523091). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

7. REFERENCES

[1] Allison, P.D. 1999. *Multiple regression: A primer*. Pine Forge Press.

[2] Baker, R.S. et al. 2012. Towards automatically detecting whether student learning is shallow. *International Conference on Intelligent Tutoring Systems* (Chania, Crete, Greece, 2012), 444–453.

[3] Baker, R.S. et al. 2012. Towards sensor-free affect detection in a Cognitive Tutor for Algebra. *Educational Data Mining* (Chania, Crete, Greece, 2012).

[4] Bixler, R. and D'Mello, S. 2016. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*. 26, 1 (2016), 33–68.

[5] Bixler, R. and D'Mello, S.K. 2015. Automatic Gaze-Based Detection of Mind Wandering with Metacognitive Awareness. *User Modeling, Adaptation and Personalization: 23rd International Conference* (Dublin, Ireland, 2015), 31–43.

[6] Bixler, R. and D'Mello, S.K. 2014. Toward fully automated person-independent detection of mind wandering. *Proceedings of the 22nd International Conference on User Modeling, Adaptation, and Personalization* (Switzerland, 2014), 37–48.

[7] Blanchard, N. et al. 2014. Automated physiological-based detection of mind wandering during learning. *Intelligent Tutoring Systems* (Honolulu, Hawaii, USA, 2014), 55–60.

[8] Boys, C.V. and others 1890. *Soap-bubbles, and the forces which mould them*. Cornell University Library.

[9] Chawla, N.V. et al. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. (2002), 321–357.

[10] Cieslak, D.A. et al. 2012. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*. 24, 1 (2012), 136–158.

[11] Dixon, W.J. and Yuen, K.K. 1974. Trimming and winsorization: A review. *Statistische Hefte*. 15, 2–3 (1974), 157–170.

[12] D'Mello, S.K. 2016. Giving Eyesight to the Blind: Towards Attention-Aware AIED. *International Journal of Artificial Intelligence in Education*. 26, (2016), 645–659.

[13] Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM*. 55, 10 (2012), 78–87.

[14] Ekman, P. and Friesen, W.V. 1977. *Facial action coding system*.

[15] Faber, M. et al. 2017. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*. (2017), 1–17.

[16] Franklin, M.S. et al. 2011. Catching the mind in flight: Using behavioral indices to detect mindless reading in real time. *Psychonomic Bulletin & Review*. 18, 5 (2011), 992–997.

[17] Franklin, M.S. et al. 2013. Window to the wandering mind: pupillometry of spontaneous thought while reading. *The Quarterly Journal of Experimental Psychology*. 66, 12 (2013), 2289–2294.

[18] Holmes, G. et al. 1994. Weka: A machine learning workbench. *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems* (1994), 357–361.

[19] Hutt, S. et al. 2016. The eyes have it: gaze-based detection of mind wandering during learning with an intelligent tutoring system. *Proceedings of the 9th International Conference on Educational Data Mining, International Educational Data Mining Society* (2016), 86–93.

[20] Jeni, L.A. et al. 2013. Facing Imbalanced Data–Recommendations for the Use of Performance Metrics. *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (2013), 245–251.

[21] Kononenko, I. 1994. Estimating attributes: analysis and extensions of RELIEF. *Machine Learning: ECML-94* (1994), 171–182.

[22] Kopp, K. et al. 2015. Influencing the occurrence of mind wandering while reading. *Consciousness and cognition*. 34, (2015), 52–62.

[23] Kopp, K. et al. 2015. Mind wandering during film comprehension: The role of prior knowledge and situational interest. *Psychonomic Bulletin & Review*. 23, 3 (2015), 842–848.

[24] Littlewort, G. et al. 2011. The computer expression recognition toolbox (CERT). *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)* (2011), 298–305.

[25] Mills, C. et al. 2016. Automatic Gaze-Based Detection of Mind Wandering during Film Viewing. *Proceedings of the 9th International Conference on Educational Data Mining* (Raleigh, NC, USA, Jun. 2016).

[26] Pham, P. and Wang, J. 2015. AttentiveLearner: Improving Mobile MOOC Learning via Implicit Heart Rate Tracking. *Artificial Intelligence in Education*. C. Conati et al., eds. Springer International Publishing. 367–376.

[27] Randall, J.G. et al. 2014. Mind-Wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological bulletin*. 140, 6 (2014), 1411.

[28] Reichle, E.D. et al. 2010. Eye movements during mindless reading. *Psychological Science*. 21, 9 (2010), 1300–1310.

[29] Risko, E.F. et al. 2013. Everyday attention: Mind wandering and computer use during lectures. *Computers & Education*. 68, (2013), 275–283.

[30] Smallwood, J. et al. 2004. Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and cognition*. 13, 4 (2004), 657–690.

[31] Wan, X. 2009. Co-training for Cross-lingual Sentiment Classification. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (Stroudsburg, PA, USA, 2009), 235–243.

[32] Webb, N. and Ferguson, M. 2010. Automatic Extraction of Cue Phrases for Cross-corpus Dialogue Act Classification. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (Stroudsburg, PA, USA, 2010), 1310–1317.

[33] Westlund, J.K. et al. 2015. Motion Tracker: Camera-Based monitoring of bodily movements using motion silhouettes. *PloS one*. 10, 6 (2015).

[34] Zacks, J.M. et al. 2010. The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in human neuroscience*. 4, 168 (2010), 1–15.

[35] Zhang, Z. et al. 2011. Unsupervised learning in cross-corpus acoustic emotion recognition. *2011 IEEE Workshop on Automatic Speech Recognition Understanding* (Dec. 2011), 523–528.

[36] 2016. *Emotient module: Facial expression emotion analysis*.