Attention and Engagement in Virtual Environments: Measuring the Unobservable

Benjamin Bell
Eduworks Corporation
Corvallis, OR
benjamin.bell@eduworks.com

Benjamin Nye
University of Southern California
Institute for Creative Technologies
Playa Vista, CA
nye@ict.usc.edu

Winston Bennett, Jr.
Air Force Research Laboratory
Wright-Patterson AFB, OH
winston.bennett@us.af.mil

Elaine Kelsey
Eduworks Corporation
Corvallis, OR
elaine.kelsey@eduworks.com

ABSTRACT

Engagement is a principal factor in learning retention. Although engagement describes behaviors (unlike motivation, which describes a cognitive state or trait), engagement is difficult to directly observe and measure in many training settings, including virtual reality (VR) simulation. Training using VR is gaining broad adoption across DoD. In the Air Force, for instance, a new approach to Specialized Undergraduate Pilot Training (SUPT) called Pilot Training Transformation (PTT), integrates traditional flying sorties with VR-enabled ground-based training devices to accelerate training. To support PTT with metrics of attention and engagement, Eduworks and USC's Institute for Creative Technologies partnered with the Air Force Research Laboratory (AFRL) to develop machine learning (ML) models that can measure user engagement during any computer-mediated training (simulation, courseware) and offer recommendations for restoring lapses in engagement. We developed and tested this approach, called the Observational Motivation and Engagement Generalized Appliance (OMEGA) in a PTT context. Two factors motivate this work. First, a goal of PTT is for an instructor pilot (IP) to simultaneously monitor multiple simulator rides. Being alerted to distraction, attention and engagement can help an IP manage multiple students at the same time, with recommendations for restoring engagement providing further instructional support. Second, the virtual environment provides a rich source of raw data that ML models can use to associate user activity with user engagement. We created a testbed for data capture to construct the ML models, based on theoretical foundations we developed previously. We ran pilots through multiple PTT scenarios and collected formative data from instructors to evaluate the utility of the recommendations OMEGA generates regarding how lapsed engagement can be restored. We present findings that validate the use of ML models for detecting engagement from the data characteristic of virtual environments. These findings, though preliminary, will support innovating and accelerating conventional and VR applications as training adapts to an unexpected future.

ABOUT THE AUTHORS

Dr. Benjamin Bell is the president of Eduworks, where he leads simulation, training, human-machine interaction, and decision support development. His research has addressed the use of simulation for training and education across a spectrum of applications, including K-12, higher education, military, and national security training. He has held faculty positions, chief executive positions in industry, and leadership roles for several international conferences. He is an adjunct professor at Embry Riddle, holds a PhD from Northwestern, and is a University of Pennsylvania graduate.

Dr. Benjamin Nye, Director of Learning Science at ICT, leads the Army's Promoting Engagement in Virtual Learning Environments (ENGAGE) project and is co-PI on the ONR Personal Assistant for Life-Long Learning (PAL3) project. His research, recognized for excellence in tutoring and behavior simulation, focuses on technologies to promote learning, and has yielded over 20 peer-reviewed papers, 12 book chapters, 1 book, and 5 open-source projects.

Dr. Winston "Wink" Bennett is the Airman Systems Directorate Readiness Product Line Lead. He is a Senior Research Psychologist in the Warfighter Interactions and Readiness Division, 711th Human Performance Wing Air Force Research Laboratory. Through more than 25 years of service in the Air Force research community, he has achieved international recognition as a leader in education, training, and performance measurement research. He has led numerous research products that have since become part of the operational military community and have significantly improved mission effectiveness. He pioneered training, education, and measurement technologies and transitioned research results to operational military, scientific and commercial communities that have produced ground-breaking training technology and research and serve as a foundation for other researchers and practitioners. He is a Fellow of three research societies and the Air Force Research Laboratory.

Elaine Kelsey is a computational linguist with Eduworks Corporation, where she focuses on natural language processing and machine learning. She led development of Eduworks' automated question generation, assessment, and competency alignment algorithms. She holds multiple bachelor's and master's degrees in computer science, linguistics, molecular biology, and biostatistics, and is working towards a MS/PhD in machine learning.

Attention and Engagement in Virtual Environments: Measuring the Unobservable

Benjamin Bell
Eduworks Corporation
Corvallis, OR
benjamin.bell@eduworks.com

Benjamin Nye
University of Southern California
Institute for Creative Technologies
Playa Vista, CA
nye@ict.usc.edu

Winston Bennett, Jr.
Air Force Research Laboratory
Wright-Patterson AFB, OH
winston.bennett@us.af.mil

Elaine Kelsey
Eduworks Corporation
Corvallis, OR
elaine.kelsey@eduworks.com

ATTENTION AND ENGAGEMENT AS RELEVANT TRAINING METRICS

Learner engagement is a key factor in both learning retention (Hu & Hui, 2012) and learning outcomes (Chi & Wylie, 2014). We use *engagement* to describe behavior – exhibiting cognitive work during an activity – and *attention* to describe deliberate perceptual focus that results from engagement, though for our purposes both terms are suitable for describing the metrics we have developed. Assessing engagement is a challenge because engagement is not a directly measurable attribute. Instructors can employ indirect measures to infer engagement levels though this can be complicated when an instructor is monitoring multiple students simultaneously, or when an instructor's view of the student is obscured by virtual reality (VR) equipment.

To support a more systematic approach to measuring engagement, Eduworks and USC's Institute for Creative Technologies developed machine learning (ML) models that can measure user engagement during training simulations and offer recommendations for restoring lapses in engagement. We embedded these models in a system, called the Observational Motivation and Engagement Generalized Appliance (OMEGA). A particular focus of our work is training in VR simulations. The virtual environment provides a rich source of raw data that machine learning models can use to associate user activity with user engagement.

We created a testbed for data capture to construct the ML models, based on theoretical foundations we developed previously (Bell, Kelsey, Nye & Bennett, 2020). Our recent research explored OMEGA's potential to help alert an instructor pilot (IP) to student distraction by flagging attention and engagement lapses. Our hypothesis is that OMEGA could help an IP adapt learning, and potentially manage multiple students at the same time, with alerts of lapsed attention and recommendations for restoring engagement.

To test this hypothesis, we ran pilots through representative scenarios to create data for training different variants of machine learning models and analyzed the performance of each using standard machine learning metrics. In this paper, we present a summary of our machine learning design and the resulting models' performance metrics, describe our formative evaluation, and present findings from this evaluation along with recommendations for the broader applicability of this work.

STUDY CONTEXT: PILOT TRAINING TRANSFORMATION

The U.S. Air Force has implemented a new approach to Specialized Undergraduate Pilot Training (SUPT) called Pilot Training Transformation (PTT), which serves as the operational setting for the OMEGA project. PTT integrates traditional flying sorties with VR-enabled ground-based training devices to achieve training efficiencies, improve readiness, and increase throughput. While PTT student pilots fly the same hours and sorties in the T-6 as do their legacy pilot training counterparts, the ground-based training is done using an immersive PC-based flight simulator. At the time of this study, PTT was using Lockheed Martin's Prepar3D®, HTC's VIVE™ Pro VR headset, stick, throttle and rudder pedals, and a syllabus of PTT scenarios (Figure 1). However due to the rapid evolution of PTT, the hardware and software used in the study does not reflect current technology used in PTT.

During scenarios flown in the simulator, objective data is readily captured for every time interval, such as aircraft state (position, attitude, airspeed) and configuration (aileron, rudder and elevator deflections; flap and gear positions). Instructors can also monitor how the scenario is progressing and can provide verbal feedback in real-time.

Some important metrics though are less directly observable or measurable. We mentioned above that student engagement is widely accepted as a critical mediating factor in both learning retention (Hu & Hui, 2012) and learning outcomes (Chi & Wylie, 2014), and the importance of engagement is not lost on instructors. Although labels like attention and focus are more common in pilot training, such metrics are often the basis for, or at least an element of, scoring situational awareness (SA).

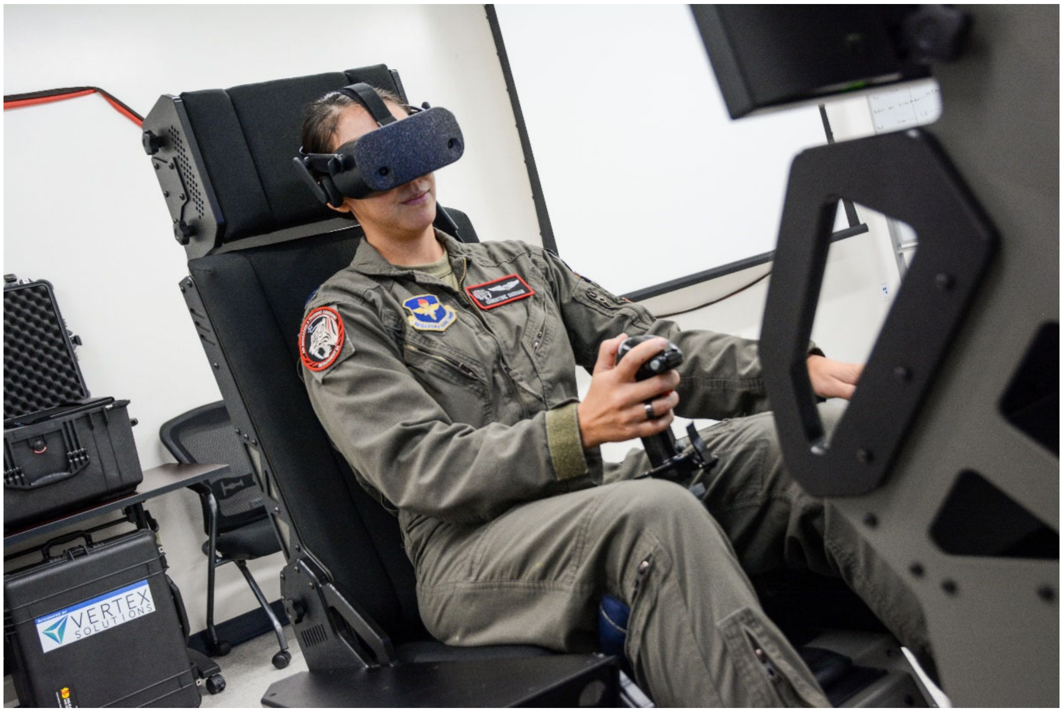


Figure 1. PTT station: VR headset, controls, displays. (Photo By: Aryn Lockhart, Air Force). The appearance of U.S. DoD visual information does not imply or constitute DoD endorsement.

Two additional factors make engagement even more salient for PTT. First, instructors have less visibility into student engagement (due to the VR headset) than in conventional simulators. Second, a vision for PTT is for one instructor to be monitoring multiple students simultaneously. Indirectly-observable measures such as engagement will thus require some level of automation support to cue instructors when lapses are detected.

MEASURING ENGAGEMENT: THE MODELING GAP

Research-based General Engagement Model

A central focus of our work has been characterizing engagement with computable measures. We started with a general model of engagement that one of us (Nye) developed from nine relevant engagement and disengagement models from the literature that emphasize behavioral indicators (e.g., data from log files or from direct queries to the user), as reported at I/ITSEC 2016 (Core, Nye, *et al.*, 2016). From this we synthesized a multi-timescale engagement and motivation model (Bell, Kelsey & Nye, 2019).

For our current effort, we refined the model to reflect the aviation focus of this project, preferring metrics associated with *event response* tasks (e.g., maneuvering to avoid a new hazard) and *monitoring* tasks (e.g., maintaining straight and level attitude). These characteristics are better aligned with how instructors monitor student pilot performance in training sorties. Maneuvers are evaluated by observing air speed, vertical speed, attitude, angle of attack, and so on. Instructors are also interested in situational awareness (SA), which they can assess by observing the student's ability to "stay ahead of the airplane": anticipating upcoming changes in heading, airspeed, or altitude; applying smooth control inputs to adjust bank angle, pitch and power; and maintaining proper scan of the flight instruments.

We thus incorporated research that identified indicators of distraction and disengagement for accidents attributed to loss of control and airplane state awareness that are relevant to flight tasks performed in simulated environments, including attention versus distraction (Harrivel, *et al.*, 2017), boredom and distraction (Cummings, *et al.,*2013), attentional tunneling (Wickens, 2005), and vigilance (Casner & Schooler, 2015). The model resulting from this additional analysis is based on eight input metrics, used to compute three mid-level features: Performance, Efficiency, and Responsiveness. An overall composite measure of current engagement within a given time window is derived from these mid-level features.

Addressing Gaps in Engagement Modeling

Engagement is an inherently subjective attribute. For PTT, a rich set of data is available from the desktop flight simulator, including aircraft position, attitude and configuration. These performance metrics are objective and can be calculated with some reliability. For instance, detecting when a student pilot lowers the gear while the airspeed exceeds the maximum gear-down speed is straightforward. To monitor engagement, however, requires aggregating observable measures to generate an indirect estimate of engagement. Our model, for instance, as mentioned above, specifies eight such indirect measures.

The addition of a VR head-mounted display (HMD) adds additional data points that could be incorporated as part of a suite of indirect metrics that could be combined to measure engagement levels. Typical VR headset and sensors can capture head position and movement; higher-end devices, such as the VIVE Pro Eye, can capture eye tracking data. This apparent abundance of data, however, does not solve the problem of developing reliable measures of engagement. Several challenges for interpreting the data remain, including, non-exhaustively: (1) Understanding which data points are relevant to engagement; (2) Setting proper coefficients representing how each data point should be weighted; (3) Distinguishing between and properly applying a single data point *x* observed at time *t* compared with a trend of how *x* behaves over some interval (*e.g.*, from *t* – 5 seconds to *t* + 5 seconds); and (4) Incorporating the velocity of the change in a data point, for instance, how abrupt an aileron deflection or throttle movement the student applied. A principal emphasis of this work is to explore the role that *machine learning models* could play in interpreting simulator and VR device data in order to develop measures to drive our conceptual model of engagement. This machine learning approach is summarized in the next section.

MACHINE LEARNING MODELS FOR MONITORING ENGAGEMENT

Model Design Approach

The intended purpose for machine learning (ML) models in this project is to learn associations between student pilot activity and predicted levels of engagement. More specifically, we employ ML to allow OMEGA to develop more accurate predictive associations between raw data inputs and higher-level aggregated engagement metrics. Our approach leverages the underlying data streams available from Prepar3D to provide better predictive power in situations where there is limited access to interpreted data (e.g. when interpreted metrics of event occurrence, event success/failure, and efficiency are not available). To achieve this, we employ three methods: The general model (our conceptual, research-based model described previously); Support Vector Machines (SVM), a standard ML technique; and Hidden Markov Models (HMM), a machine learning approach based on the statistical Markov model.

We use SVM to accurately predict engagement and disengagement in input metric sequences. This approach is attractive because it enjoys fast estimation methods with low run-time, and therefore can provide near-instant feedback to instructors. SVM techniques are most powerful in cases where sequence classification is not strongly context-dependent. For OMEGA, however, we expect some context-dependence in the data. For example, rapid adjustments of heading, altitude and airspeed may represent recovery from a period of inattention if these maneuvers occur between waypoints, but may represent an attentive reaction if observed during an event requiring active response (e.g. a heading change when passing a waypoint). To mitigate this risk and improve the model, we explored an additional ML technique more robust to sequence classification in context-dependent data.

HMM is one of the most efficient methods for modeling shorter-term dependencies between adjacent time intervals. We transformed the continuous stream of metric input data into a discrete sequence of values, based on an experimentally-validated interval of 10 seconds, which consistently produced the best results across all three methods. We used the standard Viterbi algorithm for decoding, as well as the Expectation Maximization method (specifically the Baulm-Welch Learner) to train the model using the training set time series.

Training the Models

Data for training the machine learning components came from 36 experimental subjects who each flew a pre-selected set of PTT scenarios in a data collection station that mirrors most of a PTT simulator, namely, the simulation software, stick and throttle, and VIVE Pro HMD and sensors. The data collection station also includes a dedicated application for the experimenter to monitor each scenario, interact with the subject, and time-stamp relevant events.

For purposes of creating training data for the machine learning models, experimenters are trained in a protocol to (1) time-stamp lower-intensity and higher-intensity segments of a scenario, to help the models account for workload in processing measures of user activity; and (2) engage the subject in conversation at specific points during a scenario. Conversing with the subject acts as a surrogate for disengagement. We posit that loss of attention, or distraction, will be statistically detectable in the simulation log files. Specifically, we anticipate three possible types of deviations: response time, performance, and efficiency.

Subjects were recruited from flying clubs in the Corvallis, OR and Los Angeles regions. Subjects qualified through meeting either flying hour criteria or flight simulator experience criteria. Each subject was given practice with the PTT station and then asked to complete six PTT scenarios. The collected data were used to train the ML models, comparing both the predictive power and the latency and resource requirements for each of our learning techniques.

Machine Learning Model Performance

To assess the performance of each of our models, we conducted analyses based both on the engagement/ distraction/boredom labels, and on collapsing distraction and boredom into a single proxy measure for disengagement. The datastream provided data from Prepar3D at intervals of 0.16 seconds. All results reported here are for a time interval of 10 seconds (Table 1).

Accuracy is the ratio of correctly predicted observations (boredom or distraction) to the total observations. It can be a useful measure for symmetric datasets where values of false positive and false negatives are equivalent. Since that is not the case here, we employ additional parameters to evaluate the model. *Precision* is the ratio of correctly predicted positive observations to the total predicted positive observations. In our case, this metric answers the question "of all instances of boredom the model identified, what percentage were actual instances of boredom". High precision indicates a low false positive rate. *Recall* refers to the ratio of correctly predicted positive observations to all positive observations. In this instance, Recall answers the question "of all actual instances of boredom, what percentage did the model notice?". High recall indicates a low false negative rate. The *F1 Score* is the weighted average of Precision and Recall, which takes both false positives and false negatives into account. F1 is more relevant than Accuracy in our case, because of the uneven class distribution.

HMM models consistently out-performed the other approaches we tested. Precision was higher than recall for the combined (Boredom + Distraction) model. This suggests that the model is usually correct when it detects an incident of disengagement but suffers from false negatives. In the context of OMEGA, this is a preferred balance, as we expect that instructors are more likely to trust the system if the percentage of false positives is minimized. The size of the data sets and the process for dividing the labeled data suggest that we cannot rule out overfitting as a contributing factor to the high scores. The presence of data from the same users in both the training and test sets may have produced a model that is well-tuned to the specific set of users from this trial. Collecting additional data to conduct further testing of the model on a larger range of users and scenarios would mitigate the potential influences of overfitting.

Table 1. Results of three modeling techniques.

Boredom

	Accuracy	Precision	Recall	F_1
General Model	0.744	0.497	0.550	0.522
SVM	0.759	0.523	0.580	0.550
HMM	0.825	0.630	0.750	0.685

Distraction

	Accuracy	Precision	Recall	F_1
General Model	0.760	0.382	0.530	0.444
SVM	0.770	0.404	0.550	0.466
HMM	0.823	0.508	0.700	0.589

Engagement (Boredom and Distraction Combined)

	Accuracy	Precision	Recall	F_1
General Model	0.714	0.616	0.542	0.577
SVM	0.729	0.640	0.568	0.602
HMM	0.807	0.732	0.730	0.731

RECOMMENDATION ENGINE

Adaptive Recommendation Model

OMEGA processes detect engagement levels to generate adaptive recommendations to help an instructor restore lapsed engagement. During a scenario, based on combinations of different state signals, OMEGA will generate a set of intermediate inferences. These inferences include, for example, whether poor performance is due to consistently bad results versus irregular behavior or inconsistency (e.g., carelessness). The model employs both the basic state model and the aggregated inferences as inputs to calculate a scoring ranking for different adaptive interventions.

Our model considers three levels of outcomes: performance, responsiveness, and efficiency, each representing a distinct dimension of quality. Performance represents the basic ability to complete the assigned tasks, based on the performance criteria for those tasks (e.g., following a set of waypoints). Responsiveness represents the speed and effectiveness for a learner to adjust to new tasks or requirements (e.g., if a waypoint is moved, how quickly does the user adjust heading). Efficiency represents lean and strategic use of resources to complete a scenario (e.g., faster completion times).

These quality criteria can each be thought of as building upon each other: a learner must adjust heading to a new waypoint or else there is no way to determine responsiveness. Likewise, efficiency is impossible if the user is not responsive enough to stay on course. This means that only some factors should be addressed with certain types of learners (e.g, high vs. low expertise). For example, if a user is failing to master proper take-off procedures, critiquing fuel efficiency would add no training value. On the other hand, an otherwise high-performing student pilot who is drifting off-course or leaving assigned altitudes may benefit from noting a need for improved in-flight checks.

Fig. 2. High-Level Policy for Adaptive Interventions in PTT

The policy for adaptive interventions is depicted schematically in Figure 2. The right-hand side of Figure 2 shows the interventions used for PTT. These include three distinct types of interventions: Messaging (Information about the task), Motivation (Context about the task and learning goals), and Recommendations (Suggestions on different tasks or breaks to improve learning). The left-hand side of Figure 2 outlines a high-level policy for when specific interventions are expected to be appropriate for users with different skill levels and in different states. These connections between intervention types and student states do not represent the actual model. Instead, they represent key dynamics that the model will produce. However, since the actual state space to calculate an effective intervention policy is too large to easily convert into a short graph, this model captures the key behaviors that the intervention model will be tested against, to ensure it behaves reasonably versus what would align to theoretical frameworks for engagement and responding to disengagement. The intervention types are outlined in Table 2.

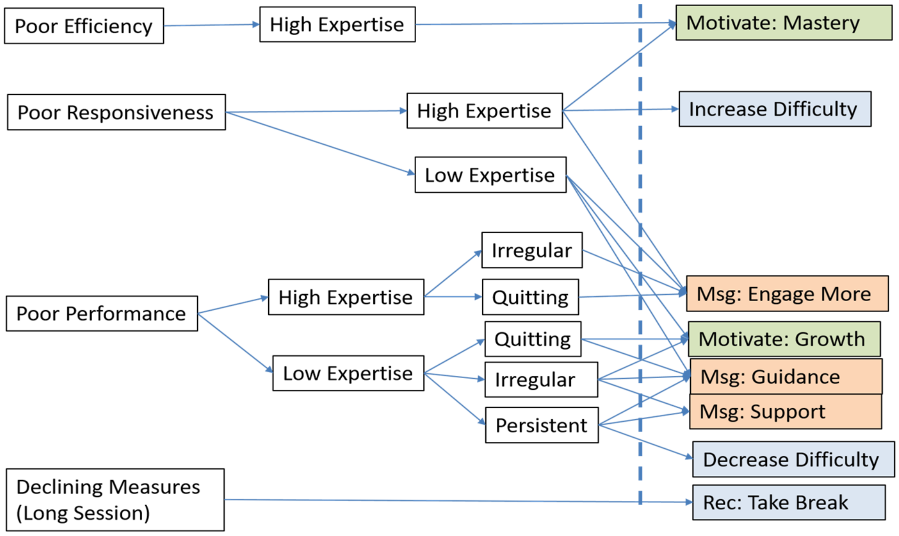


Table 2. Intervention Types for PTT

Category	Type	Description
Messaging	Engage More	A suggestion or warning that signs of decreased engagement were noticed, with a suggestion on how to improve
	Guidance	A useful hint about how to do better on the specific scenario or skills
	Support	Affective feedback to help a learner who is struggling
Motivate	Mastery	Information on how better training can improve real-life performance
	Growth Mindset	Feedback on how sustained effort leads to improved skills & outcomes
Recommend	Increase Difficulty	Suggest that user might need more challenging scenarios
	Decrease Difficulty	Suggest user might benefit from trying some easier scenarios
	Take Break	Suggest that user returns after taking a break, due to performance

Machine Learning-Driven Recommendations

We investigated two distinct methods for generating adaptation recommendations based on the internal state of the models used to detect disengagement in the sensor data stream from Prepar3D. The first approach is much less computationally-intensive and produces recommendations with lower latency. However, given the highly contextualized nature of the input data stream, we explored a second approach to produce more accurate results, and tested the trade-offs between timeliness and quality under different simulation conditions.

In the first approach, we use the calculated values from the general model of engagement (performance, efficiency, and responsiveness) and its sub-component features, as inputs into a machine learning classification model. Using the labeled data set produced during the data collection trials with licensed pilots, we applied a traditional machine learning modeling technique (SVM) to the classification task, where the outputs are the available set of adaptations available in the PTT training environment.

This technique however does not account for the contextual nature of the data stream, instead analyzing each time slice as a separate case. However, as was the case for detecting engagement levels, the determination of an appropriate adaptation depends highly on the context in which the disengagement event occurs, as well as on the environmental conditions being simulated. Here we define context as the data stream of pilot behaviors and actions preceding and following the time slice of interest, and environment as the set of conditions that obtain for that particular segment of the simulation (e.g. aircraft attitude, airspeed, vertical speed, status of systems, to name a few). To fully account for the contextual nature of both the disengagement event detection and the recommendation of an appropriate adaptation, we employed the HMM approach discussed previously.

FORMATIVE EVALUATION

To examine the utility of a tool for detecting engagement levels, we created a dashboard showing an instructor pilot (IP) the Prepar3D flight simulation view side-by-side with OMEGA's alerts triggered by lapses in engagement (Figure 4). We designed a formative evaluation with external Subject Matter Experts (SMEs) consisting of two FAA Certified Flight Instructors (25 years and 7 years of pilot-in-command flying) and one Air Force F-16 pilot and current reservist with extensive instructor experience. The small sample size was due principally to COVID-related constraints affecting access to our testbeds and to pilots. The formative evaluation focused on three aspects:

1. Recommendation Appropriateness: Rating specific recommendations as appropriate and valid;
2. Recommendation Usability: Rating if the recommendations made sense and would be useful for an instructor;
3. Diagnostic Metrics Usability: Rating if the diagnostic indicators that OMEGA uses to select among different recommendations make sense to an instructor.

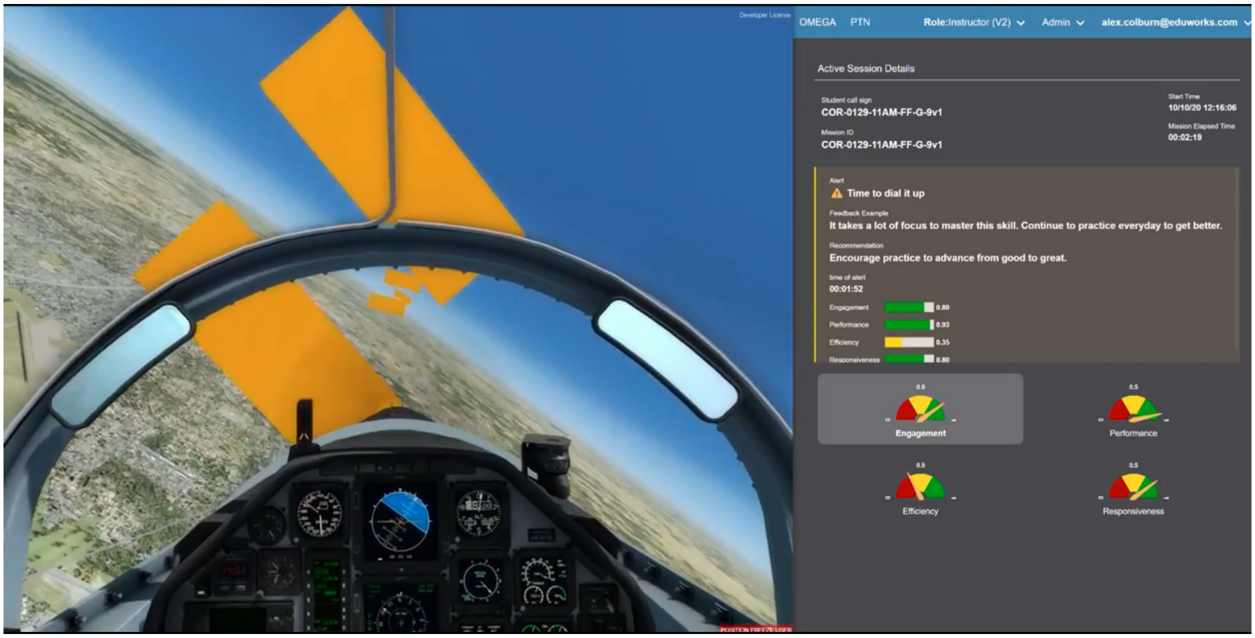
Formative Evaluation Design

To validate the recommendations produced by OMEGA, we developed a protocol for Subject Matter Experts (SMEs) to review OMEGA's recommendations, as presented in the Dashboard, side-by-side with a Prepar3D playback of a recorded flight in a PTT scenario (Figure 4). These recorded scenarios were captured during the training of the machine learning models, and as discussed previously, during these simulations, experimenters intervened to cause distraction and split attention to induce disengagement (e.g., responding to a long-answer narrative question). The scenario recordings were selected to present a variety of performance levels from a diverse cross-section of the 36 subjects.

Recommendation Appropriateness

For every recommendation given during a scenario run, the SMEs were asked to evaluate its appropriateness. These recommendations have three main components:

1. Intervention Points (time): OMEGA first attempts to recognize an area with poor flight engagement metrics, as recognized through machine learning. This is trained entirely bottom-up from data.
2. Intervention Type: Next, the system attempts to select an effective intervention at that point. This is done by analyzing expert-defined diagnostic metrics (shown to instructors as gauges in the validity task)
3. Intervention Feedback Text: The specific text that OMEGA would suggest as an example, such as: "Better, but stay ahead of the airplane." Unlike other components, these are currently not optimized and during validity ratings the SMEs were always presented with the same example text for each intervention type.



Active Session Details

Student call sign
COR-0129-11AM-FF-G-9v1

Mission ID
COR-0129-11AM-FF-G-9v1

Start Time
10/10/20 12:16:06

Mission Elapsed Time
00:02:19

Alert
Time to dial it up

Feedback Example
It takes a lot of focus to master this skill. Continue to practice everyday to get better.

Recommendation
Encourage practice to advance from good to great.

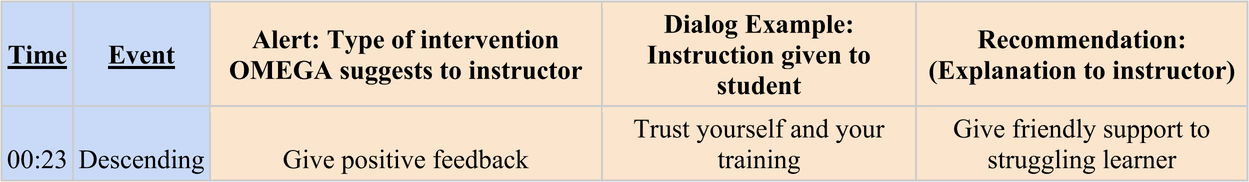
time of alert
00:01:52

Engagement	0.80
Performance	0.93
Efficiency	0.35
Responsiveness	0.80

Figure 3. Formative Assessment View: PTT Scenario (left), OMEGA dashboard (right)

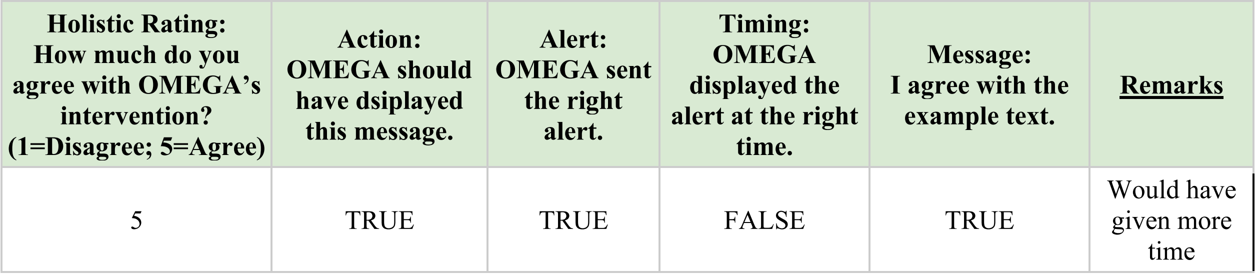
The SME validity check is primarily intended to evaluate the Intervention Points (time) and Intervention Type (what is said). Flight instructors are thus rating whether they would have spoken during that time and whether they agree with what was said. As a secondary factor, they are giving feedback on whether the example feedback is appropriate or what they might have said instead.

These ratings were performed by watching a recorded scenario with the split screen dashboard shown in Figure 3, where the experimenter would stop at each intervention point to collect these ratings (scores of 1-5 for the intervention overall, checkboxes for indicating agree/disagree with the intervention, type, timing, and feedback text). These ratings populate a spreadsheet that include both the OMEGA-generated data (Figure 4) and fields for instructor ratings (Figure 5) for each event line.



Time	**Event**	**Alert: Type of intervention OMEGA suggests to instructor**	**Dialog Example: Instruction given to student**	**Recommendation: (Explanation to instructor)**
00:23	Descending	Give positive feedback	Trust yourself and your training	Give friendly support to struggling learner

Figure 4. Rating Sheet (Pre-Populated Information).



Holistic Rating: How much do you agree with OMEGA's intervention? (1=Disagree; 5=Agree)	**Action: OMEGA should have dsiplayed this message.**	**Alert: OMEGA sent the right alert.**	**Timing: OMEGA displayed the alert at the right time.**	**Message: I agree with the example text.**	**Remarks**
5	TRUE	TRUE	FALSE	TRUE	Would have given more time

Figure 5. Rating Sheet (SME-Provided Information).

Recommendation Usability

After each scenario run, instructors were asked a short series of questions that align to traditional Technology Acceptance Model (TAM) usability metrics (Davis, Bagozzi & Warshaw, 1989):

- Ease of Use: How much SMEs understood the recommendations and found them easy to use
- Expected Benefit: How much the SMEs thought these recommendations would be useful to them as instructors, working with one or more students who practice on flight simulators

These questions were also asked about the system overall, after all scenarios were completed (Table 3).

Table 3. Example of Usability Ratings

Please rate your level of agreement from 1 to 6.	
1= Completely Disagree / 3 = Somewhat Disagree / 4 = Somewhat Agree / 6 = Completely Agree	
1. The OMEGA recommendations were easy to use and understand.	6
2. The OMEGA recommendations made sense to me.	6
3. The OMEGA recommendations were reasonable.	4
4. The OMEGA recommendations would be useful to help with training.	4
5. Any additional notes?	*Really need the feedback and altitude to make good judgements as an instructor*

Diagnostic Metrics Usability

As a final formative metric, usability ratings were also requested for the diagnostic metrics that OMEGA uses to help choose its recommendations for a specific intervention point. These metrics, as noted earlier, are Performance, Efficiency, and Responsiveness. Since the display of these metrics is not yet optimized (shown in gauges which give approximate red/yellow/green color codes for low/med/high), only a single Ease of Use question was asked for each diagnostic metric (i.e., whether the metric in the gauge made sense to the SME). These items were:

1. The Engagement gauge made sense to me.
2. The Performance gauge made sense to me.
3. The Efficiency gauge made sense to me.
4. The Responsiveness gauge made sense to me.
5. Any additional notes?

Results

Usability Ratings

In general, the validity results were generally positive but showed the need for specific improvements before using OMEGA as a live training tool. As shown in Table 4, both the Ease of Use and Expected Benefit for the tool were rated positively (as 4 or greater on a 1-6 scale). Ease of Use and Expected Benefit are averaged from two items adapted from the widely-used TAM model. Ease of Use was rated fairly high and instructors understood the elements in general. Expected Benefit as a training tool was lower, with two raters positive (4.75 and 5) and one negative (1.5). Since the qualitative feedback was similar for all three raters, we posit that this difference in ratings represents whether the SME was rating the tool primarily "as-is" or "after a few noted improvements."

Ratings of usability averaged across the individual scenarios were comparable and did not show notable trends. Ratings on the usability of the specific gauges (Engagement, Performance, Efficiency, Responsiveness) showed that some of the metrics displayed were not well understood, as shown in Table 5. Specifically, the overall Engagement gauge and Efficiency gauge were only partly understood. In some scenarios instructors felt that they agreed with these gauges, while in others they were unclear. One instructor who rated his understanding as a 3 (slight disagree) explained it as, "Engagement as I understood was some kind of combination of them all and I'll just take your word for it."

Table 4. Overall Usability Ratings (N=3).

Usability Item or Construct	Mean	StdDev
EASE OF USE	5.2	1.4
- Was easy to use and understand.	5.0	1.7
- Made sense to me.	5.3	1.2
EXPECTED BENEFIT	3.8	1.0
- Were reasonable.	3.8	1.6
- Would be useful to help with training.	3.7	2.3

Efficiency was also unclear to instructors, for two reasons. First, instructors said it was hard for them to rate efficiency themselves due to limited ability to see the flight controls and gauges. Second, in some cases they felt that pilots were flying fairly efficiently, but the system rated a flight segment's Efficiency poorly, which reduced confidence in that metric. By comparison, Responsiveness (speed to adapt to changes, such as a new waypoint) and Performance (missing waypoints, having incidents in flight) were generally well understood.

Table 5. Usability Ratings for OMEGA Gauges for Interpreting Engagement

Usability Item	Mean	StdDev
The Engagement gauge made sense to me.	3.3	2.5
The Performance gauge made sense to me.	5.7	0.6
The Efficiency gauge made sense to me.	2.7	2.1
The Responsiveness gauge made sense to me.	5.0	1.0

Recommendation Ratings

Ratings for the individual recommendations in each scenario were reasonable but showed obvious areas for improvement based on instructor feedback. Table 6 shows a summary of the average ratings for each recommendation across the three instructors. The Overall score was 2.71, on a scale of 1 (Disagree) to 5 (Completely Agree). The Should Intervene checkbox was treated by raters as an approximately binary version of this (correlation of r=0.93) and a regression indicated that each Overall recommendation score point increased Should Intervene by 0.27. This means that ratings of 3, 4, and 5 indicated that SMEs mostly would have accepted and agreed with the recommendation, while ratings of 1 and 2 mean they would have rejected it.

Table 6. Recommendation Event Ratings - Agreement Levels

Overall	Should Intervene	Type	Timing	Too Early or Too Soon
2.71	47%	40%	38%	31%

Agreement among SMEs about Overall scores did not have a high level of agreement. This was evident in the pairwise correlations between raters for Overall scores, which were 1 vs. 2 (r=0.03), 1 vs. 3 (r=0.11), and 2 vs. 3 (r=0.24). We calculated inter-rater reliability on the full event set, using a binary coding of each recommendation as accepted if the Overall rating was 3 or higher and rejected otherwise. Observed agreement was 48% between raters, with a Fleiss kappa of -0.03, over 26 recommendation events. The low kappa score appeared to be affected by redundant (Too Early or Too Soon) recommendations, where an instructor tended to pick one or the other but not both. It was also influenced by individual raters, whose preferences for feedback timing varied (e.g., some would never give feedback during an approach to a gate). As there was no clear way to consistently resolve these differences, these ratings could not be easily resolved and thus are a ceiling for the overall OMEGA validity ratings (i.e., it would be hard to surpass 50% agreement as the SMEs do not currently exceed that level). This may also in general indicate some differences in pilot instructing styles regarding timing and frequency of feedback.

When SMEs rated a recommendation well, they typically also reported that they agreed with the recommendation type (correlation of r=0.80). The most common recommendation type was either to emphasize a Push Getting Better (e.g., encourage a growth mindset where a pilot struggling with some performance issues focused on their training) or to Dial it Up (e.g., to work on smaller tweaks and changes to increase from good to great). The primary disagreement with a Type was either a Dial it Up for a pilot who was still struggling or either intervention given to a more proficient student who had not been given sufficient time to settle in.

SMEs often rated the timing as reasonable if the recommendation was reasonable (r=0.81). Cases where instructors would have intervened, but the system did not, were very infrequent (average of 3-4 total for any one instructor across all 5 scenarios). This implies that the *recall* of the system is probably fairly high. However, as noted, the volume of recommendations and their timing was overall an area that instructors emphasized as needing improvement (one instructor expressed this concern as "too much, too soon"). Overall, the most common feedback by instructors was that too many recommendations were suggested, while a flight instructor would generally "wait and see" for a longer period before giving feedback

Finally, the specific text examples for the feedback were generally reviewed as too long and too vague. Instructors preferred to use feedback that was either very short (e.g., "Eyes forward") or that was specific to the flight task (e.g., "Too much bank."). In many cases, OMEGA was detecting the existence of a problem correctly (e.g., poor Efficiency score due to overcorrecting), but is not designed to be a diagnostic model to classify these problems such that more specific feedback could be given. For the PTT scenarios, it appeared that a majority of specific issues might be possible to detect automatically based on the current data or with only minor additions (e.g., more logging of when waypoints were hit or estimating distance to waypoints). This indicates that specific feedback on PTT scenario performance might be possible using a fairly general model. An alternative approach is to harvest metrics calculated from concurrent sources, such as VIPER, which was not made available to the OMEGA team for this project.

Discussion

Our results must be interpreted as preliminary and suggestive only, due to the small sample (N=3) of instructor pilot evaluators available under COVID-related restrictions. However, these preliminary findings suggest that a major area for improvement in OMEGA is the mediation between *detection* (i.e., when a problem occurs) and *feedback* (i.e., when an instructor should intervene). Most instructors felt that OMEGA was fairly good at detecting engagement lapses and triggered recommendation alerts in response to the kinds of shaky, distracted flying that were induced during the data collection runs (e.g., the most positive instructor stated: "Could see a notification and know with 80% certainty that something should be looked at"). Based on feedback and metrics, the strengths of OMEGA are:

1. Ease of Use: The recommendations were well-understood and actionable for instructors
2. Detector Timeliness: The OMEGA classifier detected the periods when significant events indicated disengagement or poor student pilot performance, based entirely on flight data logs.
3. Recommendation Coverage: Instructors noted very few periods where they would have intervened that OMEGA missed. This is not because OMEGA was activated constantly, since in many cases it did not intervene for 45 second to over a minute, but due to a wide variety of suboptimal patterns being detected.

However, while in some cases instructors felt that immediate feedback might be appropriate, there was a significant number of conditions where instructors would either delay feedback or not give feedback at all. One instructor compared OMEGA to a new flight instructor, in that "This is what every new flight instructor does. They can't stop talking and then learn to say less." The tendency to give too much feedback is the opposite of OMEGA's detection performance, where the classifier tended to be fairly accurate but would miss certain periods of disengagement.

As a result, failing to detect a significant percentage of disengagement periods may not be a problem for the recommender. This is because the recommender already intervenes more frequently than an instructor, even though it only detects about two-thirds of the known disengagement periods. In general, the results of this formative study imply that the development of the disengagement detector has been overall quite successful, in that enough events are detected and identified in time with sufficient granularity that they can drive training recommendations.

CONCLUSION AND FUTURE WORK

The SMEs reported that substantial development is still required to optimize the recommender and feedback generation models. Comments by the SMEs showed four main deficits:

1. More Specific Feedback: "The individual messages are the part that needs to be adjusted up." Instructors noted that even if the underlying problem might be disengagement, they would not give general feedback (e.g., "Focus on your training") but would give specific feedback (e.g., "Focus on the next gate.")
2. Redundancy/Too Soon: "need to reduce quick duplicates" Instructors were universally negative on the tendency for OMEGA to give recommendations back-to-back (e.g., within less than 10 or 20 seconds). This implies that OMEGA should wait for longer periods and present a top-rated feedback message at a strategic time (e.g., after completing a training segment), rather than reactively recommending based on the current detected events.
3. Increased Flight Context: "Missing airspeed, altitude, attitude; Really need the feedback and altitude to make good judgements as an instructor." Instructors wanted to know more about what the student was doing before they would accept a message. They requested that the OMEGA dashboard show a good summary of the flight status and controls at-a-glance, so that they could determine which recommended alerts to accept.
4. Safety Alerts: "Safety parameters are the biggest glaring thing." OMEGA was designed to detect disengagement. This means that it does not detect some other factors, such as unsafe maneuvers. All instructors noted when one pilot lost altitude rapidly and stated that they would intervene immediately. As such, a separate detector should be developed to detect and react to those issues, since instructors will want to respond to them promptly, regardless of the cause.

This feedback showed several areas to improve the OMEGA system as a recommender. These areas appear actionable and with the potential to improve recommendations to a level that is adequate to assist an instructor in generating feedback messages, and eventually to assist a student directly without an instructor.

This work will help advance learning outcomes and retention for training using simulations and VR. A key factor in achieving positive outcomes is learner engagement, which is more challenging to assess than directly observable or objectively measurable factors. In some instances, the VR environment itself can obscure cues relevant to learning engagement from instructor view. OMEGA addresses this gap by using ML models to develop predictive associations between simulation events and learner actions on the one hand, and learner engagement on the other. OMEGA also incorporates a model of adaptive interventions to remedy engagement lapses, and employs ML to develop associations between the context of the engagement lapse and the optimal intervention to recommend.

REFERENCES

Bell, B., Kelsey, E., and Nye, B (2019). Monitoring Engagement and Motivation Across Learning Environments. In *Proc. of the 2019 MODSIM World Conference*. Norfolk, VA.

Bell, B., Kelsey, E., Nye, B., and Bennett, W. (2020). *Adapting Instruction by Measuring Engagement with Machine Learning in Virtual Reality Training*. In R. Sottilare & J. Schwarz (Eds.) Adaptive Instructional Systems. HCII 2020. Lecture Notes in Computer Science, vol 12214, pp. 271 – 282. Springer: Cham.

Casner, S. M., & Schooler, J. W. (2015). Vigilance impossible: Diligence, distraction, and daydreaming all lead to failures in a practical monitoring task. Consciousness and cognition, 35, 33-41.

Chi, M. T., and Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist, 49*(4), 219-243.

Core, M.G., Georgila, K., Nye, B.D., Auerbach, D., Liu, Z.F., DiNinni, R. (2016). Learning, adaptive support, student traits, and engagement in scenario-based learning. In I/ITSEC, 2016.

Cummings, M. L., Mastracchio, C., Thornburg, K. M., & Mkrtchyan, A. (2013). Boredom and distraction in multiple unmanned vehicle supervisory control. Interacting with Computers, 25(1), 34-47.

Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35* (8), 982–1003.

Harrivel, A. R., Stephens, C. L., Milletich, R. J., Heinich, C. M., Last, M. C., Napoli, N. J., & Pope, A. T. (2017). Prediction of cognitive states during flight simulation using multimodal psychophysiological sensing. In AIAA Information Systems-AIAA Infotech (p. 1135).

Hu, P. J. H., and Hui, W. (2012). Examining the role of learning engagement in technology-mediated learning and its effects on learning effectiveness and satisfaction. *Decision support systems, 53*(4), 782-792.

Wickens, C. D. (2005). Attentional tunneling and task management. In 2005 International Symposium on Aviation Psychology (p. 812).