Toward Data-Rich Models of Visitor Engagement with Multimodal Learning Analytics

Jonathan Rowe, Wookhee Min, Seung Lee, Bradford Mott, and James Lester

North Carolina State University, Raleigh, NC 27695, USA
{jprowe, wmin, sylee, bwmott, lester}@ncsu.edu

Abstract. Visitor engagement is critical to the effectiveness of informal learning environments. However, measuring visitor engagement raises significant challenges. Recent advances in multimodal learning analytics show significant promise for addressing these challenges by combining multi-channel data streams from fully-instrumented exhibit spaces with multimodal machine learning techniques to model patterns in visitor experience data. We describe initial work on the creation of a multimodal learning analytics framework for investigating visitor engagement with a game-based interactive surface exhibit for science museums called FUTURE WORLDS. The multimodal visitor analytics framework involves the collection of multichannel data streams, including facial expression, eye gaze, posture, gesture, speech, dwell time, and interaction trace log data, combined with traditional visitor study measures, such as surveys and field observations, to triangulate expressions of cognitive, affective, behavioral, and social engagement during museum-based learning. These data streams will be analyzed using machine learning techniques, with a focus on deep recurrent neural networks, to train and evaluate computational models of engagement using non-intrusive data sources as input (e.g., interaction logs, non-identifying motion tracking data). We describe distinctive opportunities and challenges inherent in using multimodal analytics within informal settings, as well as directions for utilizing multimodal visitor analytics to inform work by exhibit designers and museum educators.

Keywords: Multimodal Learning Analytics, Informal Learning Environments, Interactive Tabletop Exhibits, Game-Based Learning.

1 Introduction

Engagement is the cornerstone of learning in informal environments [1]. During free-choice learning, such as in museums and science centers, visitor engagement shapes how learners interact with exhibits, move around the exhibit space, and form attitudes, interests, and understanding of science. In recent years, the space of possibilities for visitor engagement has been enriched by the growing presence of advanced learning technologies within museums, including digital games, tangible devices, and augmented reality. Engagement with these technologies can be operationalized in many different ways, and research in the area spans a broad range of fields and theoretical perspectives [2]. Disentangling visitor engagement from related constructs, such as motivation, flow, interest, and self-regulation, is a common challenge. Engagement has

also proven difficult to measure [2]. Much of the research on learner engagement has depended upon the use of subjective measures, such as self-reports, questionnaires, and interviews. These measures provide a snapshot view of visitor engagement, but they provide limited data for modeling visitor engagement at the process level. Observational methods have also been utilized, but they raise issues of scalability, as well as potential disruptive effects inherent in video recording, audio recording, and even written consent itself [3].

Recent developments in multimodal learning analytics show particular promise for addressing these challenges. Learning analytic techniques enable the creation of computational models for inferring complex relationships between variables, which can be utilized to detect the presence of engagement-related phenomena from non-identifying data, such as trace logs of learner behavior [4]. Multimodal learning analytics expands upon these methods by using multiple physical hardware sensors to concurrently capture multi-channel data on learner behavior and modeling salient patterns of learner experience using machine learning [5, 6]. Multimodal learning analytics has shown significant promise in laboratory and classroom environments [7, 8], but there has been comparatively little work investigating multimodal learning analytics in informal contexts, such as science museums.

In this paper, we describe work on the design and development of a data-driven framework for investigating visitor engagement in science museums using multimodal learning analytics (Figure 1). We focus on visitor interactions with a game-based interactive surface exhibit about environmental sustainability called FUTURE WORLDS. By instrumenting FUTURE WORLDS with multiple hardware sensors, it is possible to capture fine-grained data on visitors' facial expression, eye gaze, posture, gesture, conversation, dwell time, and learning interactions to triangulate key components of visitors' cognitive, affective, behavioral, and social engagement during free-choice learning. We use these data streams to train and evaluate multimodal machine learning models to infer visitors' engagement levels with non-intrusive data sources as input (e.g., interaction logs, non-identifying motion tracking data). We describe recent progress on the development of the multimodal visitor analytics framework, and discuss distinctive opportunities and challenges inherent in using this approach to devise computational models of visitor engagement.

2 Background and Related Work

Learner Engagement. We adopt a conceptualization of engagement that is organized in terms of several core components: cognitive engagement, emotional engagement, behavioral engagement, and social engagement [9, 10]. Cognitive engagement describes individuals' psychological investment in learning, which has close ties to motivation and interest as well as self-regulated learning [9]. Emotional engagement refers to individuals' affective responses to learning, including attitudes, mood, and moment-to-moment emotional expressions, such as engaged concentration, delight, confusion, and surprise. Behavioral engagement refers to learners' positive, on-task, and productive learning behaviors. In a museum context, low levels of engagement may appear as passive or shallow interactions with an interactive exhibit, whereas high-level behavioral engagement can manifest as productive exploration behaviors, as well as

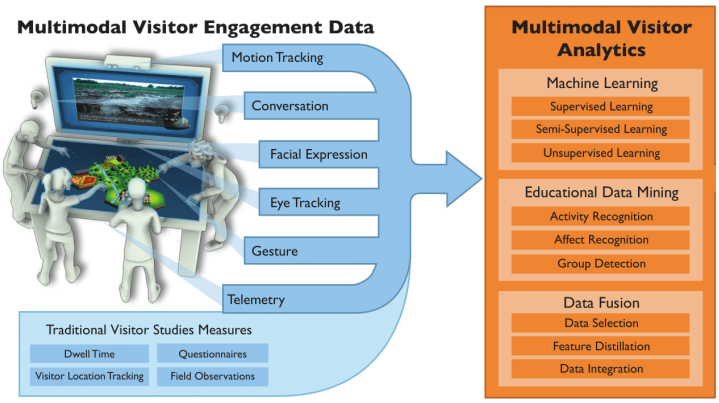

Figure 1. Multimodal visitor analytics framework.

expressions of interest that extend outside of the exhibit (e.g., prompting a friend to try the interactive tabletop display). Social engagement acknowledges the key role of social interactions during learning in small groups, a common context in museums and other informal environments [10]. Adopting this conceptualization, we seek to utilize rich multi-channel data streams to identify salient patterns of meaningful visitor engagement that integrate cognitive, affective, behavioral, and social measures.

Multimodal learning analytics. Advances in multimodal learning analytics have been enabled by the increased availability of low-cost physical sensors (e.g., motion-tracking cameras, eye trackers) combined with significant progress in machine learning tools and techniques. By taking advantage of information across concurrent sensor-based data channels, multimodal learning analytic techniques have been found, in many cases, to yield improved models compared to unimodal techniques [11]. This extends to a range of tasks within educational technologies, including automated detection of affective states [7, 12], computational models of assessment [13], and models of learner metacognition [14]. Although these applications have shown significant promise, the preponderance of work on multimodal learning analytics has been conducted in laboratory and classroom settings. Using multimodal learning analytics to investigate visitor engagement in informal environments is a natural next step for the field.

3 FUTURE WORLDS Testbed Exhibit

To conduct data-rich investigations of visitor engagement in science museums, we utilize a game-based museum exhibit called FUTURE WORLDS. Developed with the Unity game engine, FUTURE WORLDS integrates game-based learning technologies and interactive surface displays to enable collaborative explorations of environmental sustainability in science museums [15]. In FUTURE WORLDS, visitors solve sustainability problems by investigating the impacts of alternate environmental decisions on a 3D simulated environment (Figure 2). The virtual environment is rendered from a top-down perspective on a 28” interactive surface display, a Microsoft Surface Studio 2. Learners tap and swipe to test hypotheses about how different micro-

Figure 2. Screenshots from FUTURE WORLDS game-based museum exhibit.

and macro-scale environmental decisions—such as modifying a region's electricity portfolio or augmenting a farm's waste management practices—impact the environment's sustainability and future health. The effects of visitors' environmental decisions are realized in real-time with 3D game engine technologies.

FUTURE WORLDS' focus on environmental sustainability targets three major themes—water, energy (both renewable and non-renewable), and food—and it facilitates exploration of the interrelatedness of these themes. Initial pilot testing with both school and summer-camp groups at our partner museum, the North Carolina Museum of Natural Sciences, have indicated that learners' interactions with FUTURE WORLDS enhance sustainability content knowledge, as well as yield promising levels of collaboration and engagement as indicated by observations of learner behavior [15].

4 Multimodal Visitor Analytics Framework

We are investigating the use of a suite of multimodal sensors (e.g., webcam, motion-tracking camera, eye tracker, directional microphone, game logs) to capture visitors' facial expression, body movement, eye gaze, speech, and interaction trace data, respectively, to serve as complementary data sources for inducing computational models of visitor engagement with FUTURE WORLDS (Figure 3). Integration of hardware sensors with the Unity game engine is implemented with a client-server architecture, enabling multimodal data capture to be conducted on a separate thread from the main game logic for FUTURE WORLDS. This modular design supports extensibility, enabling the addition of alternate hardware sensors and exhibit software applications in the future. In addition to collecting multimodal sensor data, we utilize visitor self-reports (e.g., questionnaire data) and observational methods to obtain "ground-truth" labels of visitor engagement. These serve as the raw data for creating engagement labels, which operationalize expressions of visitors' cognitive, affective, behavioral, and social engagement for use in supervised machine learning.

4.1 Multimodal Data Channels

Facial expression. Facial expression provides a rich window into learner emotion and engagement [16]. Facial action unit data has been found to provide an effective input for recognizing learning-centered affective states [7]. In our work, we capture facial expression data using video recordings from an externally mounted Logitech C920 USB webcam. The resulting data is analyzed using OpenFace, an open-source facial behavior analysis toolkit that provides automated facial landmark detection, facial

Figure 3. Child exploring an early prototype of the FUTURE WORLDS exhibit instrumented with multimodal sensors.

action unit recognition, and eye gaze tracking functionalities [17]. OpenFace supports both real-time tracking and post-interaction analysis of facial expression, and it offers opportunities to train facial feature recognition models based on developers' needs.

OpenFace is built as a native C++ library, which is integrated with our multimodal visitor analytics server. OpenFace has demonstrated efficient run-time performance on Microsoft Surface Studio 2 hardware, tracking facial movements at approximately 30 frames per second. It also provides visualization tools for inspecting the quality of head pose and eye gaze estimation functionalities. In its initial implementation, the OpenFace integration is configured for tracking facial landmark coordinates, a single user at a time, and external webcam support, although support for multiple users is planned. Multi-face tracking comes at the expense of reduced reliability at tracking facial action units. Therefore, we plan to conduct initial studies with the single-user model and transition toward studies with multiple simultaneous visitors as the project progresses. Tracking which data belongs to which person is an engineering challenge that will be important to address as the project transitions toward more naturalistic studies in the museum with fluid grouping of visitors.

Eye gaze. Gaze provides rich, task-based information that can significantly contribute toward modeling users' cognitive and affective states [18]. A growing body of empirical work has demonstrated the importance of eye gaze for modeling learner interactions [19]. To track visitor eye gaze, we utilize a mounted eye-tracking sensor, the Tobii EyeX eye tracker, which uses near-infrared light to track eye movements and gaze points during user interactions with game environments [20]. This data is complementary to eye gaze estimation data generated by the OpenFace facial behavior analysis toolkit. We automatically identify in-game targets of user attention in FUTURE WORLDS using a gaze target-labeling module that processes eye tracking data using ray casting techniques in Unity. Using this module, it is possible to automatically track visitors' visual fixations on in-game objects and interface elements, yielding log events that contain the name of the target game object, as well as the timestamp and duration of the fixation.

Posture and Gesture. Recent years have seen growing interest in research on affective modeling using human body movement data [12, 21]. To capture data on visitor posture and gesture, we utilize Microsoft Kinect for Windows v2, a dedicated motion sensing camera that provides skeletal tracking for pose and gesture detection, as well as raw pixel data for depth and color camera sensors [22]. Currently, the software tracks visitors' pose through the use of skeletal joint orientations at a maximum rate of 30 Hz. The Kinect software can differentiate up to 6 people at a time, and it tracks 26 individual joints per person. The Kinect sensor is typically mounted on a tripod several feet away from the interactive surface exhibit, and it allows for tracking of pointing gestures and shifts in posture that characterize different behavioral signatures of visitor engagement.

Conversation. Visitor conversation is a critical form of engagement with museum exhibits. Several computational techniques have emerged to investigate speech data for affect and engagement prediction [23]. We consider two approaches: (1) investigating transcribed conversations between visitors based on automatic speech recognition (ASR) or manually generated transcriptions, and (2) utilizing acoustic feature extraction from speech data. Prominent ASR methods include hidden Markov models, Gaussian mixture models, and deep neural networks [24]. Acoustic feature extraction (e.g., spectral features, prosodic features) is supported by signal processing techniques, including Fourier transform and autocorrelation functions. We investigate open-source toolkits for speech analysis that are publicly available, such as Kaldi [25] for ASR and OpenSMILE [26] for acoustic feature extraction.

Telemetry (interaction trace logs). FUTURE WORLDS provides support for telemetry, or the generation of detailed logs of learner interactions with the digital interactive exhibit software. The log data consists of timestamped records (at the millisecond level) of visitor taps and multitouch gestures, as well as learning events and simulation states, that arise during visitor experiences. Telemetry data can be utilized to investigate how learners explore and manipulate the underlying environmental simulation provided by FUTURE WORLDS. Log data collected from visitors' learning interactions will be aligned with other data streams and analyzed. Emergent interaction patterns will be examined and coded as measures to provide insight into the dynamics of the visitor experience.

Questionnaires (self-report). We utilize several questionnaires to capture pre and post data on visitors' science content knowledge, interest, and engagement related to the FUTURE WORLDS exhibit. To examine visitor learning outcomes we will collect pre- and post assessment data using personal meaning maps and an environmental sustainability identification task. Personal meaning maps (PMMs) consist of a blank piece of paper with a brief set of instructions and a prompt phrase: sustainability. Participants use a pen to write or draw words, phrases, and pictures about their conceptualizations of the prompt phrase. The environmental sustainability identification task involves learners inspecting an illustrated picture of an environment depicting both sustainable and unsustainable environmental practices and annotating the picture by circling good practices and crossing out bad practices.

To gather data on visitors' interest in natural science, we utilize the Fascination in Science scale, an 8-item questionnaire designed for use by 10–14 year olds [27]. We also utilize the Engagement in Science Learning Activities Questionnaire [28] and a modified version of the Perceived Interest Questionnaire [29] to gather retrospective

self-reports from visitors about their engagement and interest in the FUTURE WORLDS exhibit, respectively.

Field observation methods. We utilize field observation methods to systematically collect observational data on visitors' affective states and exhibited behaviors during interactions with FUTURE WORLDS. Observation methods include (1) the Simplified Engagement Observation Protocol, and (2) the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) quantitative observational protocol. The Simplified Engagement Observation Protocol [28] is designed to score an individual's engagement in a science learning experience. It records observer impressions of cognitive, affective, and behavioral engagement of predetermined participants, and it can be used with learners of any age. Observations focus on a single participant during a fixed time period, and they are recorded on paper via Likert scale. In BROMP, a trained field observer walks around the perimeter of a study area as participants engage in a learning activity, and the field observer discreetly records holistic observations of participants' physical displays of emotion in a round robin sequence [30]. Field observers use a hand-held Android device running the HART field observation software, which enables the researcher to generate real-time timestamped codes of visitors' emotional states that most closely correspond to the participant's displayed affect at that time. Emotional states are coded in terms of discrete categories, such as engaged concentration, delight, confusion, boredom, surprise, and frustration.

4.2 Deep Recurrent Network-Based Data Fusion

There are a broad range of well-established machine learning techniques that have been widely used in the learning analytics community. These include supervised learning algorithms (e.g., J48 decision trees, random forests, support vector machines, logistic regression), unsupervised learning techniques (e.g., k-means clustering, expectation-maximization), and methods that can account for the sequential nature of visitors' learning interactions, such as hidden Markov models and dynamic Bayesian networks. Many of these techniques have tradeoffs among one another, and the choice of machine learning technique often depends upon the relevant task and dataset.

In recent years, deep neural network-based methods have shown particular promise in multimodal machine learning applications [11]. To induce multimodal machine learning models of visitor engagement, we are investigating two fusion method-based long short-term memory recurrent neural network architectures (LSTMs): an early fusion-based LSTM model and a late fusion-based LSTM model. LSTMs are a variant of recurrent neural networks that are specifically designed for sequence labeling [31]. LSTMs have achieved high predictive performance in various sequence labeling tasks, often outperforming standard recurrent neural networks by preserving a long-term memory and effectively addressing the vanishing gradient problem [32].

Figure 4A shows an early fusion-based LSTM model in which joint information across all modalities per time step is represented in a shared representation space and is used as input for the model to predict visitor engagement. For early fusion-based LSTMs, it is important to conduct an explicit alignment of multi-channel data streams so that data from different modalities is appropriately synchronized. Figure 4B shows a late fusion-based engagement model, in which each modality is managed by a

separate LSTM, and the resulting modality-based models' outputs are concatenated into a summary representation, which is used to infer visitor engagement levels. For both data fusion methods, non-sequential data (e.g., self-reported traits, such as gender and age, or attitudes, such as science interest) are treated separately; they are directly linked to the output layer of the multimodal learning analytics model.

It is notable that a fusion-method LSTM framework is scalable, in principle, from individuals to groups of visitors. One approach is to stack individual visitors' models, as illustrated in Figures 4A and 4B, such that the output layers of individual visitor models serve as intermediate layers to infer group-level engagement in a late-fusion manner. Alternatively, in cases where sub-groups are fluidly formed during visitor interactions with FUTURE WORLDS, individual visitor models associated with each sub-group can be stacked together to account for sub-group formations, and then the sub-group models can be hierarchically configured to compose a full-group model.

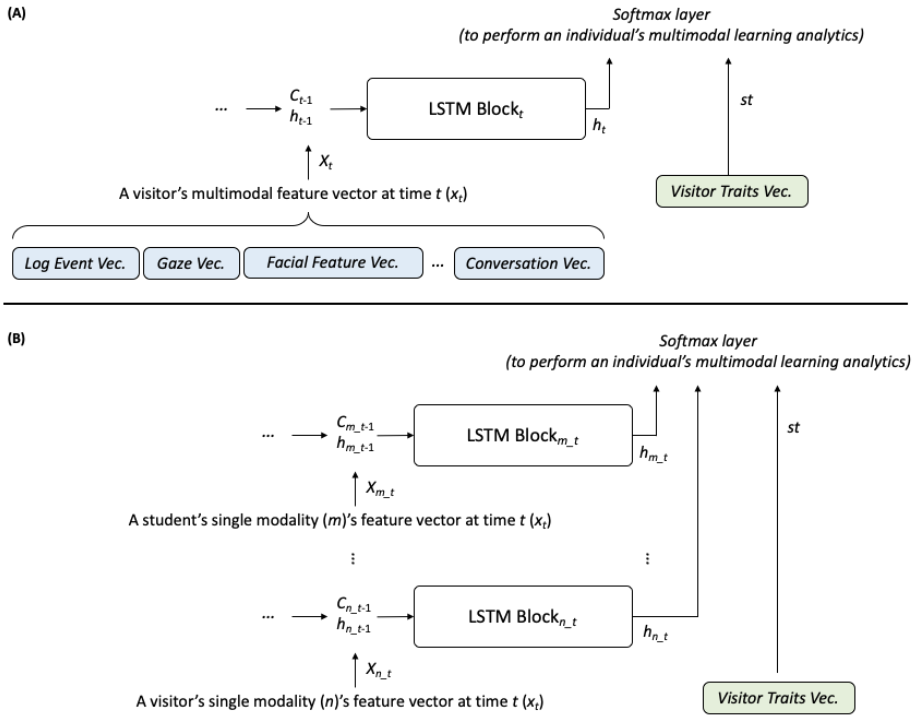

Figure 4. Data fusion-based LSTM recurrent neural network architectures: (A) Early fusion-based LSTM and (B) Late fusion-based LSTM.

5 Opportunities and Challenges

A significant opportunity afforded by applying multimodal learning analytics to investigate visitor engagement in science museums is to inform the best practices of exhibit designers and museum educators, as well as inform the design of practitioner-

focused learning analytic tools. For example, multimodal visitor analytics could be utilized to yield models that provide real-time analytics to inform how museum educators are allocated across an exhibition space to enhance high-quality visitor engagement on a busy day. These analytics could reveal to what extent meaningful engagement occurs when the museum is crowded, or at different times of the day, as well as the dynamics of visitor engagement in the presence or absence of large school groups. Multimodal analytics for visitor engagement also hold promise for informing iterative cycles of design and development by exhibit designers.

Despite significant promise, research on multimodal visitor analytics also raises challenges that merit attention. First, there are a range of open questions in multimodal machine learning about how to best address challenges centered around data representation, data alignment, data fusion, and co-learning [11]. Data representation is a challenge about how to encode multi-channel data streams in a machine-interpretable vector space, effectively representing, summarizing, and exploiting complementary information provided by heterogeneous modalities. Data alignment involves identifying dependencies between different modalities. For example, a visitor's gesture or facial expression may have been caused by another visitor's previous utterance or behavior; it is important to identify relations between events captured across different modalities. Finally, data fusion and co-learning address questions about how to join information from multiple modalities into a single predictive model, how to leverage the predictive capacity given by each modality, and how to deal with different levels of noise occurring across different modalities. Addressing these challenges will be critical for devising accurate and reliable multimodal learning analytic models of visitor engagement.

A second challenge is gracefully handling the different kinds of noise (e.g., missing data, incorrect data) that naturally arise in many sensor-intensive data collections. A range of computational methods have been devised to handle missing data, including EM imputation, temporal belief-based imputation, and multiple imputation [33]. We will investigate these techniques in connection to missing data issues in our own work on multimodal learning analytics. Devising best practices for the positioning, configuration, and calibration of physical hardware sensors in museum spaces are also likely to be necessary. For example, in eye tracking, the position and the angle of the eye tracker, the operating distance between the visitor's eyes and the eye tracker, and the nature of nearby light sources all may need adjustment to ensure high-quality eye tracking. Depending on the visitor population, it can be important for sensors to support easy mounting and position adjustments to fit individual visitors. Recommended Tobii EyeX operating distance between a user's eye and the device is between 450mm and 800mm (Gibaldi et al., 2017), and researchers should ensure that the operating distance is properly managed. Lastly, since Tobii EyeX uses infrared light, too much light or direct sun light source should be taken into account to avoid negative impacts on eye tracking accuracy.

A third challenge, which has begun to receive growing attention in recent years, is related to artificial intelligence (AI) ethics, including the risk of encoding implicit bias within machine learning-based models of visitor experience [34]. In general, machine learning models reflect the data that they have been trained on; if systematic biases exist in the training data, then similar biases are likely to arise in the machine learning models, too. This points toward the importance of recruiting diverse groups of visitors

to participate in studies that yield datasets for training and evaluating machine learning models, as well as utilizing coding schemes and measures that are culturally sensitive; prioritizing sensitivity to cultural and demographic differences between visitors is important to avoid miscategorizing or neglecting different expressions of engagement. A related challenge is preservation of learner privacy. Collecting, managing, storing, and modeling multimodal data from museum visitors in a manner that respects individuals' privacy will be essential if multimodal visitor analytics is to eventually transition from research to practice.

6 Conclusion

Multimodal visitor analytics offers significant potential to enrich our understanding of visitor engagement during free-choice learning in informal environments. By utilizing learning analytics to recognize patterns within and between dimensions of visitor engagement that are reflected across vast amounts of multimodal data, we aim to devise a rich empirical account of meaningful visitor engagement among individual visitors and small groups, as well as uncover broader tidal patterns in visitor engagement that unfold across the exhibit space. We are utilizing a suite of multimodal sensors, including eye trackers, motion-tracking cameras, webcams, microphones, and interaction trace logs, as well as traditional visitor studies measures such as questionnaires and observational protocols, to construct a data-rich account of visitors' cognitive, affective, behavioral, and social engagement during museum studies conducted with the FUTURE WORLDS game-based museum exhibit. The resulting datasets will serve as the raw input to multimodal machine learning and educational data mining techniques, and in particular data fusion-based LSTM recurrent neural networks, to induce models for classifying visitor engagement levels from concurrent complementary data streams.

Acknowledgments

We would like to thank our collaborators at the North Carolina Museum of Natural Sciences, James Minogue from North Carolina State University, and Cathy Ringstaff from WestEd for their contributions to this research. This material is based upon work supported by the National Science Foundation under grant DRL-1713545. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Bell, P., Lewenstein, B., Shouse, A. W., Feder, M. A.: Learning science in informal environments: People, places, and pursuits 32(3), 127. Washington, DC: National Academies Press (2009).
2. Sinatra, G. M., Heddy, B. C., Lombardi, D.: The challenges of defining and measuring student engagement in science. Educational Psychologist 50(1), 1–13 (2015).

3. Block, F., Hammerman, J., Horn, M., Spiegel, A., Christiansen, J., Phillips, B., Diamond, J., Evans, M., Shen, C.: Fluid Grouping: Quantifying Group Engagement around Interactive Tabletop Exhibits in the Wild. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 867–876. ACM (2015).
4. Baker, R.S., Siemens, G.: Educational data mining and learning analytics. Sawyer, K. (Ed.) Cambridge Handbook of the Learning Sciences: 2nd Edition, 253–274 (2014).
5. Blikstein, P., Worsley, M.: Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. Journal of Learning Analytics 3(2), 220–238 (2016).
6. Oviatt, S., Grafsgaard, J., Chen, L., Ochoa, X.: Multimodal learning analytics: Assessing learners' mental state during the process of learning. In: The Handbook of Multimodal-Multisensor Interfaces, pp. 331–374. Association for Computing Machinery and Morgan & Claypool (2018).
7. Bosch, N., D'Mello, S. K., Baker, R. S., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., Zhao, W.: Detecting student emotions in computer-enabled classrooms. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, pp. 4125–4129. (2016).
8. DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., Baker, R. S., Lester, J. C.: Detecting and addressing frustration in a serious game for military training. International Journal of Artificial Intelligence in Education 28(2), 152–193 (2018).
9. Fredricks, J. A., Blumenfeld, P. C., Paris, A. H.: School engagement: Potential of the concept, state of the evidence. Review of educational research 74(1), 59–109 (2004).
10. Linnenbrink-Garcia, L., Rogat, T. K., Koskey, K. L.: Affect and engagement during small group instruction. Contemporary Educational Psychology 36(1), 13–24 (2011).
11. Baltrušaitis, T., Ahuja, C., Morency, L. P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2), 423–443 (2018).
12. Henderson, N., Rowe, J., Mott, B., Brawner, K., Baker, R., Lester, J.: 4D Affect Detection: Improving Frustration Detection in Game-Based Learning with Posture-Based Temporal Data Fusion. In: Proceedings of the 20th International Conference on Artificial Intelligence in Education, Chicago, Illinois (2019).
13. Emerson, A., Sawyer, R., Azevedo, R., Lester, J.: Gaze-Enhanced Student Modeling for Game-based Learning. In: Proceedings of the 26th ACM Conference on User Modeling, Adaptation and Personalization, pp. 63–72, Singapore (2018).
14. Taub, M., Azevedo, R.: Using Sequence Mining to Analyze Metacognitive Monitoring and Scientific Inquiry based on Levels of Efficiency and Emotions during Game-Based Learning. JEDM Journal of Educational Data Mining 10(3), 1–26 (2018).
15. Rowe, J. P., Lobene, E. V., Mott, B. W., Lester, J. C.: Play in the Museum: Design and Development of a Game-Based Learning Exhibit for Informal Science Education. International Journal of Gaming and Computer-Mediated Simulations (IJGCMS) 9(3), 96–113 (2017).
16. Monkaresi, H., Bosch, P. Calvo, R., D'Mello, S. K.: Automated Detection of Engagement using Video-Based Estimation of Facial Expressions and Heart Rate. IEEE Transactions on Affective Computing 8(1), 15–28 (2016).
17. Baltrušaitis, T., Robinson, P., Morency, L. P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016).
18. Min, W., Mott, B., Rowe, J., Taylor, R., Wiebe, E., Boyer, K. E., Lester, J.: Multimodal goal recognition in open-world digital games. In: Proceedings of the 13th Artificial Intelligence and Interactive Digital Entertainment Conference (2017).
19. Aung, A. M., Ramakrishnan, A., Whitehill, J. R.: Who are they looking at? Automatic Eye Gaze Following for Classroom Observation Video Analysis (2018).

20. Gibaldi, A., Vanegas, M., Bex, P. J., Maiello, G.: Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. Behavior research methods 49(3), 923–946 (2017).
21. Patwardhan, A., Knapp, G.: Multimodal Affect Recognition using Kinect. arXiv preprint arXiv:1607.02652 (2016).
22. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE multimedia, 19(2), 4–10 (2012).
23. Dhall, A., Kaur, A., Goecke, R., Gedeon, T.: Audio-video, student engagement and group-level affect prediction. In: 20th Proceedings of the International Conference on Multimodal Interaction, pp. 653–656. ACM (2018).
24. Graves, A., Mohamed, A. R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, pp. 6645–6649. IEEE (2013).
25. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. IEEE Signal Processing Society (2011).
26. Eyben, F., Wöllmer, M., Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia, pp. 1459–1462. ACM (2010).
27. Chung, J., Cannady, M. A., Schunn, C., Dorph, R., Bathgate, M.: Measures Technical Brief: Fascination in Science. Retrieved from: http://www.activationlab.org/wp-content/uploads/2016/02/Fascination-Report-3.2- 20160331.pdf (2016).
28. Chung, J., Cannady, M. A., Schunn, C., Dorph, R., Bathgate, M.: Measures Technical Brief: Engagement in Science Learning Activities. Retrieved from: http://www.activationlab.org/wp-content/uploads/2016/02/Engagement-Report-3.1-20160331.pdf (2016).
29. Schraw, G.: Situational interest in literary text. Contemporary Educational Psychology 22(4), 436–456 (1997).
30. Ocumpaugh, J., Baker, R., Rodrigo M.: Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences, 60 (2015).
31. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735-1780 (1997).
32. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks 61, 85–117 (2015).
33. Kim, Y. J., Chi, M.: Temporal Belief Memory: Imputing Missing Data during RNN Training. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 2326–2332. (2018).
34. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York, Crown Publishers (2016).