Detecting Wheel-spinning and Productive Persistence in Educational Games

V. Elizabeth Owen[1], Marie-Helene Roy[1], K. P. Thai[1], Vesper Burnett[1], Daniel Jacobs[1], Eric Keylor[1], Ryan S. Baker[2]

[1]Age of Learning, 101 N. Brand Blvd, Glendale CA 91023
[2]Graduate School of Education, University of Pennsylvania, 3700 Walnut St., Philadelphia, PA 19104
v.elizabeth.owen@gmail.com, marie.roy@aofl.com, kpthai@gmail.com, vesper.burnett@aofl.com, daniel.jacobs@aofl.com, eric.keylor@aofl.com, ryanshaunbaker@gmail.com

ABSTRACT

Games in service of learning are uniquely positioned to offer immersive, interactive educational experiences. Well-designed games build challenge through a series of well-ordered problems or activities, in which perseverance is key for working through in-game failure and increasing game difficulty. Indeed, persistence through challenges during learning is beneficial not just in games but in other contexts as well, with grit and perseverance positively associated with academic performance and learning outcomes. However, recent studies suggest that not all persistence is positive, suggesting that many students end up "wheel-spinning", spending considerable time on a topic without achieving mastery. Thus, it is vital to differentiate productive and unproductive persistence in order to understand emergent student progress, particularly in the context of learning games and personalized learning systems, in which individual pathways differ greatly based on student needs. Leveraging Educational Data Mining methods, this study builds a detector of wheel-spinning behavior (differentiated from productive persistence) in an adaptive, game-based learning system. With the ability to predict unproductive persistence early, this detection model can be used to intelligently adapt to students needing further support in-system, as well as informing in-person intervention in a classroom setting—thus supporting a personalized, engaging learning experience in both formal and informal learning environments.

Keywords

Behavior detection, predictive modeling, productive persistence, wheel-spinning, educational games, personalized learning

1. INTRODUCTION

Games as learning vehicles can offer engaging, interactive experiences in which the player has agency in exploring and solving well-ordered problems or challenges in a learner-responsive environment [1, 2]. Well-designed games seamlessly embed meaningful instruction in authentic, narrative-driven learning contexts (with the potential to assess learning in the natural progression of play [3]). As such, they have the ability to optimize learner motivation and learning trajectories without removing the experience of personal discovery [4]. By nature, games encourage discovery of an underlying rule system through boundary testing, making experimentation and failure a core part of play progression [5]. In this sense, moving through in-game failure and challenge with perseverance can be fundamental to the experience of learning in games (e.g. [6, 7]). Hence, games offer a particularly relevant context for productive persistence or grit—the ability to steadily maintain an action or complete a task despite failure or adversity (cf. [8]). Indeed, keeping players in a "flow" state of persistence [9] through a series of challenges of increasing difficulty is key to the design of "good" games, particularly in educational contexts [10]. Recent research suggests these qualities in games support student growth in areas such as academic learning, socio-emotional skills, and creative problem solving (e.g. [11, 12, 13, 14]).

Indeed, persevering through challenges during learning is beneficial not just in games but in other contexts as well. From undergraduates to military cadets to Spelling Bee competitors, findings suggest that persistence forecasts strong performance in rigorous, achievement-based learning contexts [15]. In many cases, persistence is also associated with academic achievement [16], creativity [17], and long-term outcomes like earnings and later schooling [18].

However, recent research suggests that not all persistence is positive. "Wheel-spinning" is a form of unproductive effort, where students spend too much time struggling to learn a topic without achieving mastery [19]. Wheel-spinning behaviors have been associated with reduced motivation [20] and avoiding asking for help when needed [21]. In fact, recent empirical investigation has demonstrated that wheel-spinning can be differentiated from productive persistence in an intelligent tutoring system, in real-time, determining during problem-solving whether a student's persistence will be productive [22]. Making this type of differentiation could also be valuable in learning games contexts. Persistence is important in games just as in other settings [23], with evidence suggesting that persisting unproductively in games can be a highly frustrating experience (e.g. [24]). Since challenge and problem solving are often core components of learning experiences, particularly in game-based environments, it becomes increasingly important to differentiate productive persistence (e.g. grit) from unproductive persistence (e.g. wheel-spinning) in the context of play. This differentiation could be used to offer different pathways to students based on real-time performance.

There is evidence that this type of modeling is feasible; related game-based research has shown that the same surface behavior in games can have different meanings, with distinction of productive vs unproductive failure in a games context (cf. [25]).

In this study, we empirically investigate wheel-spinning vs productive persistence in an adaptive, game-based learning system for early childhood math skills called *Mastering Math*. Specifically, we use predictive analytics to infer whether a student is engaging in wheel-spinning or productive persistence in *Mastering Math*. This detection model can be used to intelligently adapt to students needing further support in-system, as well as informing in-person intervention in a classroom setting—thus supporting a personalized learning experience in both formal and informal learning environments.

2. METHODS AND DATA COLLECTION

2.1 Game-based Learning Content

Mastering Math (MM) is a game-based adaptive learning system designed to help elementary age children build a strong understanding of fundamental number sense and operations, ranging from counting to 10 to adding and subtracting three-digit numbers using the standard algorithm. The app constitutes approximately 130 games, covering number sense and operations concepts and skills for pre-kindergarten through second grade. Each individual game maps to a learning objective, and is supported by an interactive instruction level, as well as several layers of scaffolding and feedback. In addition, the game system as a whole uses cohesive narrative and interactive characters (embedded at the level of individual games) to support student engagement with the learning world. Adaptivity functions within individual games to provide scaffolding with each level of skill difficulty, between games to adjust to students' difficulty needs, and across the system to give players a customized pathway between skills based on performance. Assessment is embedded throughout the play experience, including game-based pretests and final assessment tasks at a granular skill level.

2.2 Experimental Design

In the fall of 2018, two research studies were conducted to evaluate the effectiveness of *MM* in preschool (Study 1) and kindergarten (Study 2) students. Students in both studies came from ethnically diverse, low-income, public school districts in Southern California.

Both studies employed a cluster-randomized trial design, in which half of the participating classrooms in each study were randomly assigned to use the *MM* app as part of their classroom instruction (treatment group), while the other half used business-as-usual mathematics instruction and materials (control group). The treatment group students (394 students in total, 146 from Study 1, 248 from Study 2) were asked to use *MM* in small group settings for 15 minutes per day for three days per week, over a total of 12 weeks. After classroom implementation, overall usage averaged 5.6 hours in Study 1, and 5.22 hours in Study 2. Both treatment and control groups received a paper-and-pencil standardized assessment of early mathematics performance before and after the implementation of *MM*.

2.3 Event-stream Data Collection

Event stream data were collected using a learning game data framework based on ADAGE (Assessment Data Aggregator for Game Environments; [26]), focusing on key learning mechanic milestones as context for performance information and results, as well as comprehensive coverage of player interaction and system feedback (e.g. [27]). These milestones are called *units*, and represent repeating progress mechanics through the learning game. Generally, students can play many games in the system (each of which corresponds with a mathematics skill), and each game contains multiple levels of difficulty. Thus, a larger unit of play is a game, and within a game a student can play one level (or activity) at a time. Each activity is built to support and assess knowledge of an individual math skill. All unit starts and ends are marked in the data, and all player interactions, system feedback, and results are recorded in the *context* of the active units at the time of the event. For example: if a student taps on the screen, we capture the basic x,y coordinate, the object being tapped (if applicable), and the units that were active during the tap (e.g. which game, activity and round the player was in when the event fired). In-game performance information (e.g. *result* or score) is embedded at the unit level, recorded at the end of applicable rounds and activities. In terms of raw *player interaction*, data collected consists of taps and drags. System feedback, also called *system event*s, consists mainly of the game communicating with the player in giving formative feedback. This includes tutorial prompts, instructional input, and inactivity prompts (given if students have not interacted with the screen in 30 seconds). Additionally, every log file event is seeded with metadata such as student ID, timestamp, and session ID. Data structured in this fashion (Figure 1) allows for a comprehensive event-stream record that is labeled consistently across the system—which currently contains over 130 activities—all aligned with learning design for interpretability, a key element of viable data use for feature engineering and analysis.

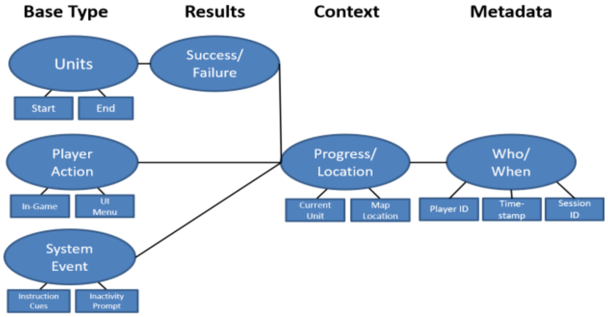


Figure 1. A simplified view of *MM's* log file data schema.

2.4 Behavior Detection

To investigate player patterns of wheel-spinning in *Mastering Math*, prediction modeling was used to build a behavior detector (i.e. model of student behavior), an automated model that can infer from log files whether a student is behaving in a certain way (e.g. [28]). These models can be employed to detect a variety of important aspects of the learner and his/her performance, including student learning, strategy, and engagement (e.g. [29, 12, 30, 31]). To train the predictive model, detectors often leverage human judgment of student behavior, in a process where behavior labels derived from human judgement are used to train and validate models, which can then automatically detect the target behavior in the larger event-stream. In this case, once the initial student interaction with a digital learning environment is captured, the analysis process includes: 1) distilling data features potentially relevant to the behavior construct; 2) identifying instances of the behavior through human evaluation; and 3) predictive modeling with the synchronized log file data.

Throughout these phases of analysis, a critical element of the data mining approach is emphasis on the event-stream trajectories that emerge in relationship to the behavior. With this detector study, each student's event-stream play patterns were observed and coded individually for emergent wheel-spinning behavior. Specific actions and click-stream interactions then emerged as evidence of wheel-spinning through the prediction algorithm’s variable selection processes. Thus, player choices and interactions characteristic of wheel-spinning were derived from the larger event-stream data flow in the analysis process detailed below.

2.4.1 Feature Distillation

In this analysis, data features were distilled from *MM* event-stream data based on play across the entire system, then refined along themes of progression and performance. These organizing themes help capture student trajectories across the system for behavior detection, particularly since student progress and failure/success are central to the target constructs of wheel-spinning and productive persistence.

Using the learning progress mechanics, or units (i.e. games or activities) from the event-stream data schema, data features were organized based on performance within each unit, as well as measures of progression (e.g. time elapsed, number of activities completed, number of games activated, etc.). (For reference, when a game is activated, it means that a student failed the associated pretest and that gameplay for that skill is now open.) Summary features were also created in parallel to the unit features, giving a sense of the overall trajectory of the player through the learning space. Since PreK and K students are in developing stages of cognition, additional features were engineered to represent age and elements of motor skill (e.g. miss rate, or how often a student drags an object towards a target and misses). One view of selected event-stream features is given below (Table 1).

Table 1. Overview of selected event-stream features

	Progression	*Performance*
Overall	• total duration in system (active play) • total activities completed • total activities started • total games started • miss rate • student age	• % of skills mastered • ratio of “boss” activities successfully passed* • total # of skills (games) activated • total skills mastered (games completed)
Game	• game completion rate • avg duration to game completion • # of answers submitted per player per game • # of activities completed within each game • avg time elapsed between activities in the same game	• individual game status: - in-progress - passed game (skill completed) - struggling (fail states for 3 of the last 5 activities) - not started - pretest passed** - pretest failed • % of started games successfully completed
Activity	• activity completion rate • avg activity duration • # of hints given • # of inactivity prompts • # of tutorials accessed	• score • progression to next level (pass/fail) • # of rounds passed • # of rounds failed • # of rounds completed

*The “boss” activity is the most difficult assessment in a game

**Pretests are embedded at the game level to test prior knowledge

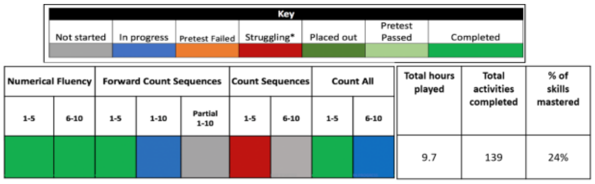
2.4.2 Behavior Coding of Wheel-spinning

For behavior detection, we focused on the construct of wheel-spinning, since the ability to flag this particular behavior held strong utility for enabling automated scaffolding in-system as well as in-person teacher intervention. Wheel-spinning is also an especially relevant focus for a game context—a medium in which boundary testing is an implicit norm [25], and differentiating real struggle from more productive forms of exploration and self-paced discovery can be valuable. *Mastering Math* games are sufficiently different from the intelligent tutoring systems, where wheel-spinning was initially studied, to require a different operationalization of wheel-spinning. In this context, we view wheel-spinning as connected to lower gameplay efficiency in the system, since wheel-spinning occurs when a great deal of effort yields very little progress [19]. To capture efficiency in an adaptive games context, in which every student has a different learning pathway, we designed a metric allowing efficiency to be standardized across players. This measure of learning efficiency was called *rate of mastery*, designed to measure the rate at which students were mastering math skills. This was calculated as the number of boss activities (the hardest assessment level in each math skill game) a student passed, divided by his/her total number of activities. This measure made sense as a progress-based metric, since performance on boss-level skill assessments is central to learning game progression. This ratio was ultimately calculated using data from both school studies. In the main behavior analysis, in accordance with the focus on wheel-spinning students and those persevering through difficulty, we concentrated on students in the lower two quartiles of *rate of mastery*.

As noted in Kai et al., 2018, we cannot assume that all lower efficiency students in the system are hopelessly struggling—on the contrary. Students who take their time to learn material, use self-paced progression, and achieve eventual success are likely to be demonstrating productive persistence. Determining whether a low-efficiency student is spinning their wheels or persisting productively is challenging. To differentiate these two groups, we started by leveraging human judgement on a per-student level to capture emergent patterns in the data. In particular, we chose to utilize the human capacity for pattern recognition and behavior evaluation (rather than an a priori rule-based approach), since the system is adaptive and no two students are likely to have the same path through the learning space.

Thus, the next step was to have human researchers observe a stream of student actions and identify the student’s behavior (e.g. [32]). The human evaluation of student behavior establishes when the behavior occurred (which serves as the predicted variable). For coding of wheel-spinning behavior in this study, play visualization based on text replays were adopted for their efficiency and accuracy [33]. Text replays, based on recorded log file data, are a text-based representation of student action during a given period of time. Text replays have shown to be highly time-efficient and scalable [50], and almost as accurate for detecting student behavior as other methods such as live observation [34].

The variation on the text replay that we used—called a visual progress replay (VPR)—includes color coding of performance levels (in addition to text summaries) for greater ease of information processing (cf. [35]). This approach represents the same information as a text replay, but in a form that encodes information with color consistent with canonical visualization techniques [36], and has previously been used to create detection models in related game-based learning research (e.g. [25]). Key features of the VPR included a visual display of game status, per student, across the system. This color-coded visual map showed whether each game was in-progress, completed (passed), in a struggle state (i.e. at least 3 non-passing scores within the last 5 activities), or not started. In addition, summary statistics per student were shown, such as number of activities completed in the system and time spent in total (Figure 2).



Key						
Not started	In progress	Pretest Failed	Struggling*	Placed out	Pretest Passed	Completed
⬜	🟦	🟧	🟥	🟩	🟩	🟩

Numerical Fluency		Forward Count Sequences			Count Sequences		Count All		Total hours played	Total activities completed	% of skills mastered
1-5	6-10	1-5	1-10	Partial 1-10	1-5	6-10	1-5	6-10			
🟩	🟩	🟩	🟦	⬜	🟥	⬜	🟩	🟦	9.7	139	24%

Figure 2. A sample portion of a VPR used for coding wheel-spinning, shown for a single student across the system.

We designed the replay's clip size to show one student's full system playthrough at a time, since we wanted to be able to detect a system-wide wheel-spinning state for each child. To capture the full trajectory of play, coding was done at the student level, labeling each student at the end of the study (week 12), in terms of whether a student was WS (wheel-spinning), P (productive persistence), and NA (not enough information). Within the lower efficiency group of students, WS captured a state of high effort but little progress, P fit with steady student progress, and NA was applied when there wasn't enough information (e.g. not enough time or activities in game to make a judgement). We included NA in this schema so that we could *derive* time and activity minimums for WS vs P differentiation through the predictor itself, rather than picking an arbitrary cutoff in excluding student data (such as, for instance, dropping all students who played for under 30 minutes). This third code also allowed for more nuanced coding—rather than forcing all students to fit under WS or P, thus risking miscategorization, the NA code could be used instead. Using this tri-code schema, inter-rater reliability analysis yielded a Cohen's κ [37] of .78, indicating acceptable agreement between raters was achieved.

2.4.3 Modeling Early Detection of Wheel-spinning

The final predictive model merged the initial feature engineering of event-stream features with the behavioral codes generated in the analysis above. To support early intervention for in-system personalization as well as teacher interventions, the final model was built to predict wheel-spinning (WS) at the end of week 12 (the last week of the study) using predictors from week 4 data. (Week 4 predictors were selected after subsequent weeks 5 and 6 were tested for model performance, but resulted in only marginal improvement.) Since each classroom was assigned exactly 12 weeks of play relative to start date, weeks as a time marker helped consistently align student progress across classrooms in relationship to the study design. It also allowed for implementation-focused behavior detection for the highest utility to teachers. With earlier detection of students getting stuck in the system, intervention can have greater impact on student progress in building core math skills.

Ultimately, the log file features (Table 1) were used as predictors in the model, while the behavior of wheel-spinning became the predicted variable. Using this full feature list, the WS detector was then built at the student level using RStudio, using the RWeka package for data mining [38]. An appropriate set of algorithms were selected based on the categorical dependent variable, informed by related behavior modeling research in education (e.g. [39, 40]), including J48, CART, Random Forest, and Naïve Bayes. Models were evaluated using ten-fold cross validation, with a final selection based on the goodness metric of AUC ROC.

To achieve higher accuracy in correctly detecting and classifying the target class of students, the wheel-spinning students, we used rebalanced classes for all three methods tested. This approach is in alignment with similar detector-based analyses in digital learning contexts (e.g. [41, 42]). Specifically, the original classes P, WS and NA had a respective number of instances of 79, 39 and 14. We set the target number of instances for each rebalanced class to n=100. To obtain that number of observations per class, for a total of n=300, sampling with replacement was performed on each class. This resampling procedure was only performed on the training set for tree building purposes and all testing was performed on the original data distribution.

3. RESULTS AND DISCUSSION

3.1 Results

Ultimately, the CART algorithm produced the best model performance, achieving a cross-validated AUC of .676 in predicting wheel-spinning in week 12 with week 4 predictors, comparable to metrics in other game-based learning detector models (e.g. [39, 12]).

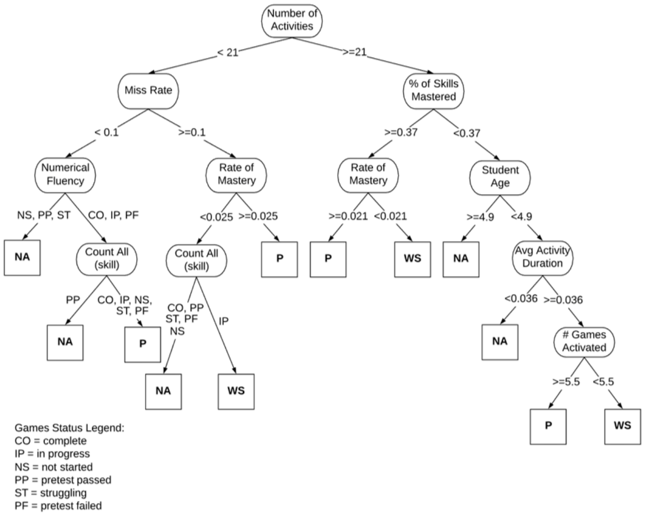


Figure 3. Final CART wheel-spinning predictor model.

Interestingly, in Figure 3 the first decision on the tree is the number of activities completed, with many students having less than 21 activities categorized as NA (insufficient information). Under this parent node, three core pathways emerge for wheel-spinners, covered from left to right on the tree: 1) low prior knowledge and low motor skill (via count all game, a low-level skill, and miss rate); 2) higher % of skills mastered but with very low efficiency (rate of mastery); and 3) younger students with low prior knowledge and very low efficiency in the system (with few games activated, higher time per activity and high number of activities per game). The first path suggests a group of students that may be in an earlier stage of development, both in terms of motor skill and prior knowledge (but not necessarily age). The second group, with the ability to master more skills with seemingly fewer motor skill issues, may represent students that have more advanced development and prior knowledge but may need a bit more scaffolding and just-in-time support to learn the material. The last group may consist of younger students with low prior knowledge who need support more directly related to age-based maturity levels. These groups, implying differing levels of development and age, reflect clinical research which suggests wide variation in the relationship between age and developmental level in young children [43]. This suggests that developmental stage (rather than age alone) is a helpful differentiator in personalizing learning experiences for young students. To investigate implications of age and development for better learning design, these emergent groups suggest value in deeper exploration of student profiles in future work (discussed below). Overall, these model-derived groups offer insight into potential types of wheel-spinning that occur within the system, with the tree model allowing for early detection of unproductive persistence.

3.2 Discussion and Conclusion

Overall, the model of wheel-spinning yields insight into the important differentiation between unproductive and productive persistence, revealing multiple ways that student wheel-spinning manifests in data and enabling event-stream detection of this behavior in the event stream data. In turn, this real-time prediction can allow for very early intervention—both in-system and in classroom—for students displaying wheel-spinning behavior. For the system, this means it may be possible to offer more intelligent adaption to student needs, while for teachers (with limited time and resources) it may become possible to offer just-in-time information about which students most need help. This emergent behavior detection is especially important in games, which can have unexpected player pathways due to complex elements of narrative, agency, and failure-driven exploration—all of which converge to support the medium's power of engagement in well-designed playful learning experiences.

Along this line of research in future work, there is an opportunity to generalize this detector to children using the game-based learning system outside of a study-specific context. The week-level data used in this study was centered around implementation, designed to flag to a teacher which students might be wheel-spinning after a certain amount of prescribed weekly dosage; however, converting this progress marker to an activity/elapsed time-based unit to build a model based on data in the wild can make this model applicable to an even broader base of learners. In addition, comparison between the classroom-based and broader event-stream based models may yield interesting insights. There is also an opportunity in this rich data stream (currently thousands of students) to hone the model for even higher AUC and predictive power. This includes iteration in feature engineering based on patterns that may arise in the larger data stream of students, using predictive modeling of wheel-spinning in a broader context of students (in formal and informal learning environments). Investigating player profiles based on detection results may also help determine groups of students struggling with the system based on motor skill, prior knowledge, and age/grade. Relatedly, better understanding how motor skill indicators in the data connected to more traditional measures of visual-motor skill (e.g. [44]) may also be valuable. Finally, dashboards highlighting detector-based insights to both parents and students for interpersonal support represent a key area for future work, with potential for student-level flagging for intervention, specific skills needing support (see Figure 2), recommendations for in-person follow-up, and possible grouping of students in the same class for differentiated instruction.

Future work in expanding the scope of wheel-spinning research in the *MM* system can support the ability to generalize findings across broader age ranges and geographic areas, increasing the potential for impact on data-driven design, intelligent personalization, and interpersonal intervention. With information on behaviors like wheel-spinning and productive persistence, in combination with other evidence such as student prior knowledge, this work can inform designers about which instructional design in games needs revisiting, as well as providing adaptive logic and system overlays for just-in-time detection and intervention. Both in the system and beyond, this research can further the application of educational data mining to principled learning design, potentially expanding the field of intelligent game-based learning and supporting young learners in developing foundational academic skills at scale.

4. ACKNOWLEDGMENTS

We would like to thank the Mastery team at Age of Learning, including Doug Dohring, Sunil Gunderia, and Diana Hughes.

5. REFERENCES

[1] J. P. Gee, "Learning by design: Good video games as learning machines," *E-Learn. Digit. Media*, vol. 2, no. 1, pp. 5–16, 2005.

[2] K. Squire, "From content to context: Videogames as designed experience," *Educ. Res.*, vol. 35, no. 8, pp. 19–29, 2006.

[3] V. J. Shute, "Stealth assessment in computer-based games to support learning," in *Computer Games and Instruction*, S. Tobias and J. D. Fletcher, Eds. Charlotte, NC: IAP, 2011, pp. 503–524.

[4] L. P. Rieber, "Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games.," *Educ. Technol. Res. Dev.*, vol. 44, no. 2, pp. 43–58, 1996.

[5] K. Salen and E. Zimmerman, *Rules of play: Game design fundamentals.* Cambridge, MA: MIT Press, 2004.

[6] K. Bielaczyc and M. Kapur, "Playing epistemic games in science and mathematics classrooms," 2010.

[7] V. J. Shute *et al.*, "Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game," *Comput. Educ.*, vol. 86, pp. 224–235, Aug. 2015.

[8] A. L. Duckworth and P. D. Quinn, "Development and Validation of the Short Grit Scale (Grit–S)," *J. Pers. Assess.*, vol. 91, no. 2, pp. 166–174, Feb. 2009.

[9] M. Csikszentmihalyi, *Flow: The psychology of optical experience*. New York: Harper Perennial, 1990.

[10] J. P. Gee, "Big 'G' Games," *http://www.jamespaulgee.com/node/63*, 2012. .

[11] J. Hamari, D. J. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke, and T. Edwards, "Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning," *Comput. Hum. Behav.*, vol. 54, pp. 170–179, Jan. 2016.

[12] R. S. Baker and J. Clarke-Midura, "Predicting Successful Inquiry Learning in a Virtual Performance Assessment for Science," in *Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization*, New York, NY, 2013, pp. 203–214.

[13] T. R. A. Kral *et al.*, “Neural correlates of video game empathy training in adolescents: a randomized trial,” *Npj Sci. Learn.*, vol. 3, no. 1, p. 13, Aug. 2018.

[14] C. Steinkuehler and S. Duncan, “Scientific Habits of Mind in Virtual Worlds,” *J. Sci. Educ. Technol.*, vol. 17, no. 6, pp. 530–543, Sep. 2008.

[15] A. L. Duckworth, C. Peterson, M. D. Matthews, and D. R. Kelly, “Grit: Perseverance and passion for long-term goals.,” *J. Pers. Soc. Psychol.*, vol. 92, no. 6, pp. 1087–1101, 2007.

[16] A. E. Poropat, “A meta-analysis of the five-factor model of personality and academic performance.,” *Psychol. Bull.*, vol. 135, no. 2, pp. 322–338, 2009.

[17] V. Prabhu, C. Sutton, and W. Sauser, *Creativity and Certain Personality Traits: Understanding the Mediating Effect of Intrinsic Motivation*, vol. 20. 2008.

[18] J. Deke and J. Haimson, “Valuing Student Competencies: Which Ones Predict Postsecondary Educational Attainment and Earnings, and for Whom?,” p. 150.

[19] J. E. Beck and Y. Gong, “Wheel-Spinning: Students Who Fail to Master a Skill,” in *Artificial Intelligence in Education*, vol. 7926. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 431–440.

[20] G. Sedek and M. Kofta, “When cognitive exertion does not yield cognitive gain: Toward an informational explanation of learned helplessness.,” *J. Pers. Soc. Psychol.*, vol. 58, no. 4, pp. 729–743, 1990.

[21] J. T. Dillon, *Questioning and Teaching: A Manual of Practice.* New York: Teachers College, 1988.

[22] S. Kai, M. V. Almeda, R. S. Baker, C. Heffernan, and N. Heffernan, “Decision Tree Modeling of Wheel- Spinning and Productive Persistence in Skill Builders,” *J. Educ. Data Min.*, vol. 10, no. 1, pp. 36–71, 2018.

[23] K. E. DiCerbo, “Game-Based Assessment of Persistence,” *Educ. Technol. Soc.*, vol. 17, no. 1, pp. 17–28, 2014.

[24] V. E. Owen, D. Ramirez, A. Salmon, and R. Halverson, “Capturing Learner Trajectories in Educational Games through ADAGE (Assessment Data Aggregator for Game Environments): A Click-Stream Data Framework for Assessment of Learning in Play,” presented at AERA, Apr-2014.

[25] V. E. Owen, G. Anton, and R. S. Baker, “Modeling User Exploration and Boundary Testing in Digital Learning Games,” in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, New York, NY, 2016, pp. 301–302.

[26] R. Halverson and V. E. Owen, “Game Based Assessment: An Integrated Model for Capturing Evidence of Learning in Play,” *Int. J. Learn. Technol. Spec. Issue Game-Based Learn.*, vol. 9, no. 2, pp. 111–138, 2014.

[27] V. E. Owen and R. S. Baker, “Fueling Prediction of Player Decisions: Foundations of Feature Engineering for Optimized Behavior Modeling in Serious Games,” *Technol. Knowl. Learn.*, vol. 24, pp. 1–26, 2018.

[28] R. S. Baker, A. T. Corbett, and K. R. Koedinger, “Detecting student misuse of intelligent tutoring systems,” in *Intelligent tutoring systems*, 2004, pp. 531–540.

[29] M. A. Sao Pedro, R. S. Baker, J. D. Gobert, O. Montalvo, and A. Nakama, “Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill,” in *User Modeling and User-Adapted Interaction*, 2013, pp. 1–39.

[30] K. E. DiCerbo and K. Kidwai, “Detecting Player Goals from Game Log Files,” in *Proceedings of the 6th International Conference on Educational Data Mining*, Massachusetts, USA, 2013, pp. 314–316.

[31] J. Asbell-Clarke, E. Rowe, and E. Sylvan, “Assessment design for emergent game-based learning,” in *CHI’13 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, 2013, pp. 679–684.

[32] R. S. Baker and A. de Carvalho, “Labeling student behavior faster and more precisely with text replays,” in *Proceedings of the 1st International Conference on Educational Data Mining*, Massachusetts, USA, 2008, pp. 38–47.

[33] R. S. Baker, “Mining data for student models,” in *Advances in intelligent tutoring systems*, Springer, 2010, pp. 323–337.

[34] R. S. Baker, A. T. Corbett, and A. Z. Wagner, “Human classification of low-fidelity replays of student actions,” in *Intelligent Tutoring Systems*, New York, NY, 2006, pp. 29–36.

[35] J. A. Wise *et al.*, “Visualizing the non-visual: Spatial analysis and interaction with information from text documents,” in *Information Visualization*, 1995, pp. 51–58.

[36] E. R. Tufte and P. R. Graves-Morris, *The visual display of quantitative information*, vol. 2. Cheshire, CT: Graphics press, 1983.

[37] J. Cohen, “A coefficient of agreement for nominal scales.,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[38] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[39] S. Kai, M. V. Almeda, R. S. Baker, N. Schectman, C. Heffernan, and N. Heffernan, “Modeling Wheel-spinning and Productive Persistence in Skill Builders,” in *Proceedings of the 10th International Conference on Educational Data Mining*, Massachusetts, USA, 2017.

[40] M. Wixon, R. S. Baker, J. D. Gobert, J. Ocumpaugh, and M. Bachmann, “WTF? detecting students who are conducting inquiry without thinking fastidiously,” in *User Modeling, Adaptation, and Personalization*, Springer, 2012, pp. 286–296.

[41] N. Bosch, Y. Chen, and S. D’Mello, “It’s Written on Your Face: Detecting Affective States from Facial Expressions while Learning Computer Programming,” in *Intelligent Tutoring Systems*, vol. 8474. Cham: Springer International Publishing, 2014, pp. 39–44.

[42] S. Kai *et al.*, “A Comparison of Video-Based and Interaction-Based Affect Detectors in Physics Playground.,” in *Proceedings of the 8th International Conference on Educational Data Mining*, Massachusetts, USA, 2015, pp. 77–84.

[43] J. Squires, D. Bricker, and L. Potter, “Revision of a Parent-Completed Developmental Screening Tool: Ages and Stages Questionnaires,” *J. Pediatr. Psychol.*, vol. 22, no. 3, pp. 313–328, 1997.

[44] K. E. Beery, “The Beery-Buktenica developmental test of visual-motor integration.” NCS Pearson., Minneapolis, MN, 2004.