Predicting Executive Functions in a Learning Game: Accuracy and Reaction Time

Jing Zhang[1], Teresa Ober[2], Yang Jiang[3], Jan Plass[1] and Bruce Homer[4]
[1]CREATE Lab, New York University, New York, NY 10012
[2]LAMBS Lab, University of Notre Dame, Notre Dame, IN 46556
[3]Educational Testing Service, Princeton, NJ 08540
[4]The Graduate Center, City University of New York, New York, NY 10016
jz1220@nyu.edu, tober@nd.edu, yjiang002@ets.org, jp79@nyu.edu, BHomer@gc.cuny.edu

ABSTRACT

Executive functions (EF) are a set of psychological constructs defined as goal-directed cognitive processes. Traditional EF tests are reliable, but they are not able to detect EF in real-time. They cause a test effect if implemented multiple times. In contrast, learning games have the potential to obtain a real-time, unobtrusive measurement of EF. In this study, we analyzed log data collected from a game designed to train the EF sub-skill of shifting. We engineered theory-based game-level and level-specific features from log data. Using these features, we built prediction models with students' accuracy and reaction time during play to predict their standard measure of the EF shifting skill during the post-test and delayed post-test as well as to predict learning gains. Our model that predicts the post score has a correlation of 0.322 and that for the delayed post score is 0.303. The findings suggest that theory-based feature engineering and varying levels of granularity are two promising directions for cognitive skills prediction under the goal of game-based assessment. Also, accuracy, reaction time, and player progression are important features.

Keywords

Prediction, Game-Based Assessment, Learning Games, Executive Functions, Cognitive Skills.

1. INTRODUCTION

Executive functions (EF) are defined as "cognitive processes used for effortful, controlled, and goal-directed thinking and behavior" [29, 3, 4]. The unity/diversity model [24] views EF as consisting of related yet separable skills, which include updating, shifting (also termed cognitive flexibility), and inhibition. EF plays an important role in cognitive development and is associated with academic success [6], metacognitive skills [7], science learning [15], and language acquisition skills [10].

Game-based assessments allow educators to assess students' learning while they are playing a game and thus in a manner that can be highly efficient, fast, and entirely unobtrusive. Using games as assessments creates a context in which learners are likely to be highly engaged, which may optimally reflect their abilities [16, 28]. Using log data from digital games to evaluate learning is sometimes referred to as a "stealth assessment" [20] and has been used in the past decade to assess complex skills, such as creativity [33] and problem-solving, [34] based on log data. Log data collected during gameplay provides a record of student behaviors associated with EF and can be used for the prediction of EF [25]; however, is it possible to use log data collected from a game designed to train EF to measure EF *and* to develop a framework for game-based assessment?

Past studies of game-based assessments have focused on complex thinking skills, such as problem-solving [34]; however, there are constraints of game-based assessments of EF. First, it is necessary to determine the granularity or time scale for which we can detect students' EF in log data. Second, we need to separate log data related to EF training from log data related to other aspects of play to achieve a high performance of models. Third, we need to generate theory-based features relevant to EF skills. Accuracy and reaction time have been identified as indicators of EF [8].

This paper aims to provide proof of the concept for game-based assessments of the EF sub-skill of shifting. Shifting is one dimension of EF defined as the ability to switch attention between different "tasks, operations, or mental sets" [21]. The research questions include:

1. How do students' gameplay data predict their executive functions during a post-test and a delayed post-test?
2. Which features, including accuracy or reaction time, are important for predicting EF in games?

2. RELATED WORK

2.1 Games for EF Training and Measurement

Sustained and active engagement is widely thought to be critical for cognitive skills training games to be effective [2]. Incorporating gamified design features is one well-established mechanism for promoting meaningful engagement [9].

Digital training games can not only enhance EF [22, 27] but can also be used as a reliable means for measuring EF and other cognitive skills. For example, past work has examined the design and validation of computerized tools for measuring working memory capacity [21]. Previous research has also examined the use of a digital game for the detection of executive functions validated by a task for medical purposes in older adults using computational modeling [12]. Further work is needed to validate game-based measures of cognitive skills [35], especially those that are sensitive enough to detect variations among neurotypical individuals and that are appropriate for child and adolescent populations.

2.2 Game-Based Assessment

In previous research, the analysis of log data as a means of a formative assessment has yielded promising findings [11] and has been used for predicting a variety of cognitive and behavioral constructs, including quitting [19], knowledge [1], computational thinking [30], persistence [26], and implicit learning [31].

Evidence-Centered Design (ECD) [23] has been used effectively to develop game-based assessments in contexts that teach specific knowledge domains [14]. According to ECD, an assessment framework should take these models into account:

- *Task model:* Which actions is the learner taking within the system?
- *Evidence model:* Which features (e.g., from log data) can be used as evidence of learner actions?
- *Competency model:* How are these features associated with a set of standards or criteria that demonstrate effective learning has taken place?

Accounting for these three ECD models is helpful for feature engineering and predictive modeling. In game-based learning contexts that teach knowledge and skills, connecting a task model to an evidence model should be relatively straightforward given that the log data provides a detailed record of the learner's action sequence. Unlike game-based assessments of knowledge domains, where standards are clearly defined and may be validated by an expert review of content, in cognitive skills training game-based contexts, further work must be done to align an evidence model with a competency model. Accuracy and reaction time have been identified as two major aspects of an EF measurement [24], with evidence suggesting they each contributes uniquely to EF performance among children [8] and adolescents [5]. Yet, the way to distinguish the nuanced forms of accuracy and reaction time at varying levels of granularity and the way to combine them for predictive modeling within a game are currently unclear.

3. DATASET

3.1 Game Design

All You Can E.T. (AYCET) [17] is a game that trains the EF sub-skill of shifting. Its early prototype, The Alien Game, has been shown to improve EF after 1.5 hours of play for high school students [16] and two hours for college students [27]. In the current study, we used the "hot" version of AYCET, a version that maximizes the playfulness of the game. As Figure 1 shows, a player is asked to feed aliens with the appropriate food based on a certain rule. The rule changes multiple times at each level, thereby requiring the player to shift. As the player progresses in the game, the rule becomes more complicated.

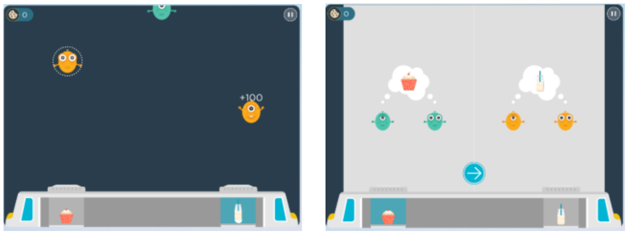

Figure 1. Feeding aliens and instructions for a rule.

3.2 Participants

Participants were recruited from three middle schools and two high schools in urban school districts in the Northeastern United States. They completed the study during non-instructional time at their schools. Among the 448 students who consented, 137 students were strategically randomly assigned to one of the three conditions to play AYCET throughout the study. Of those, 56 were removed because they demonstrated off-task behaviors, and thus the log data could not reflect their true ability. This resulted in an analytic sample of 81. Details of participant removal are discussed in the Data Cleaning subsection.

The 81 participants (M_{age} = 13.9 years, SD_{age} = 1.6, 46.1% female) included 39 in grade 7, 18 in grade 8, and 21 in grade 9. They reported a culturally and linguistically diverse background. Among them, 51.3% reported speaking Spanish at home, while 47.4% reported English and 1.3% Mandarin. As for ethnicity, 78.2% were Hispanic/Latino, 1.3% were Asian, 17.8% reported two or more ethnicities, 1.3% reported another ethnicity but did not specify, and 1.3% did not know. A few participants did not report their demographic information.

3.3 Study Procedure and Data Collection

The four-week intervention was conducted at the participating schools. Before gameplay, students completed a pretest. Then, they played the game for four sessions, each of which took about 30-40 minutes. The cumulative amount of time of play was 2-3 hours. After play, students completed a post-test, and an additional 4-8 weeks later, they completed a delayed post-test. The EF sub-skill of shifting was measured by the Dimensional Change Card Sorting (DCCS) task [36] in the pretest, post-test, and delayed post-test.

The log data consisted of 144,187 data points or actions, recording whether students fed each target ("alien") correctly or not and the reaction time for each target. This means that each alien required one action from the student. In this study, students played levels 1-30. There are 30-80 aliens per level.

In this study, each session began a few levels back from the last level played. After a few sessions, students were mandatorily pushed to level 11 to ensure they had enough time to play more difficult levels. This affected 72% of the students who were at level 9 or lower at the moment. On average, they were pushed by 4.3 levels.

3.4 DCCS Test and Score

The DCCS task [36] was used to measure the EF sub-skill of shifting in the pretest, post-test, and delayed post-test. Scoring was based on the National Institute for Health (NIH) scoring procedure [37]. This is a combination of the accuracy score and the reaction time score. The score ranges from 0 to 10. Floor or ceiling effects were not observed with our participants, as the top 25% of pretest scores ranged between 7.78 and 9.36.

4. METHOD

4.1 Data Cleaning

Based on the researchers' observations, we removed 33 participants for the following reasons: (1) did not complete one of the DCCS tests, (2) were off-task during the DCCS test, or (3) experienced technical difficulties that would affect their performance in the DCCS test. Furthermore, 23 participants were removed due to off-task behaviors (e.g., sleeping, non-stop talking, etc.) or an absence for at least one intervention session.

Eighty-one students remained in the analytic sample. The retention rate was 59.1%, which is acceptable for two reasons. First, data collected in the classroom setting are usually messier than that in the lab setting. During the study, a few participants went to the bathroom for a long time, which would be less likely to happen in a lab setting. Second, the game's focus on training EF required a degree of attention that some students were not willing to invest. Some students found it difficult to remain attentive for an extended period of time.

4.2 Labels

Table 1 lists the labels for prediction. The post score and delayed post score were directly measured by the DCCS test. We next calculated the post-learning gain and delayed post-learning gain.

Table 1. Labels

Name	Description
post score	The EF score for the post-test.
post-learning gain	Relative gain of the EF score for the post-test compared with the pretest. Based on Hake's formula of learning gain [13], it is calculated as (post score - pre score)/(10 - pre score) because the EF score ranges from 0 to 10.
delayed post score	The EF score for the delayed post-test.
delayed post-learning gain	Relative gain of the EF score in the delayed post-test compared with the pretest.

4.3 Feature Engineering

We generated 20 game-level features and five level-specific features for each level that indicated student performance and progress. They capture information related to accuracy and reaction time in various mathematical formats and granularities.

Level-specific features were features for a single level. They included the average reaction time, the standard deviation of reaction time, accuracy, the number of correct hits (i.e., an action of feeding an alien with the correct food), and the number of wrong hits (i.e., an action of feeding an alien with the wrong food) across all aliens in a single level. Accuracy was calculated as the number of correct hits divided by the total number of aliens in a level.

Game-level features were aggregated features across all levels. They included: (1) the average, maximum, minimum, range, and standard deviation of a student's accuracy across all levels after calculating the accuracy for a single level across all aliens in that level; (2) the average, maximum, minimum, range, and standard deviation of a student's reaction time across all levels after taking the average reaction time for a single level across all aliens in that level; (3) the total number of correct hits (82% of all aliens among all students), wrong hits (16%), and missed hits (i.e., an action that the student did not feed an alien) (1%); (4) the highest number of stars a student received across all levels and the total number of stars a student received in the game; (5) the number of levels a student skipped by choice (which only happened before level 10) and due to the mandatory push; and (6) the highest level and the total number of levels a student played (as a student may skip a few levels).

4.4 Model Training

We used the linear regression for predictive modelling in RapidMiner 9.3. We evaluated the model's performance using ten-fold cross-validation at the student level to ensure the model would be generalizable to a new student population. During this process, students were randomly split into 10 groups. For each possible combination, we used forward selection to select features and then built the model based on the training data. Forward selection was an iterative process. First, a single-feature model that would achieve the highest Pearson correlation was chosen. Next, the remaining features were subsequently added one-by-one to the model if they could appreciably improve the model goodness of fit. In addition, to avoid collinearity, we set the minimum tolerance for eliminating collinear features as 0.05 and set "eliminate collinear features" as true in the linear regression operator.

In addition, we explored different combinations of features for feature selection. Missing values existed for many level-specific features. Thus, we began with the first feature set containing all game-level features and level-specific features of the level with the smallest missing value rate. After that, in each round, we added level-specific features of another level based on the ranking of the missing value rate. We stopped doing so at a level that contained missing data for 16% of students. The last model contained 65 features. In this way, we controlled for over-fitting and ensured the models were trained on representative levels.

5. RESULTS

5.1 Intervention Effect

The paired samples t-test show that the post score (mean = 6.98, SD = 1.56) is significantly higher than the pretest score (mean = 6.31, SD = 2.12) ($t(80) = 3.01$, $p < 0.01$, *Cohen's d* = 0.34). Also, the delayed post score (mean = 7.16, SD = 1.33) is significantly higher than the pretest score ($t(80) = 3.90$, $p < 0.001$, *Cohen's d* = 0.43). The boxplot of three EF scores is shown in Figure 2.

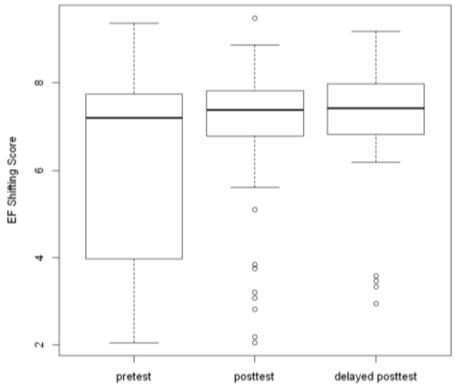

Figure 2. Boxplot of three EF shifting scores.

5.2 Correlation between Features and Labels

Among the 65 features for modeling training, five features had an absolute value of correlation with the post score between 0.3 to 0.42. Eight features had an absolute value of correlation with the delayed post score between 0.3 to 0.45. Features were weakly correlated with two learning gain labels as the absolute values of all correlation coefficients are below 0.25. Missing values for each feature were replaced with the average value of that feature.

5.3 Findings from Predictive Models

Cross-validated metrics for the best models of each label and their features are summarized in Table 2.

Table 2. Summary of the predictive models

	Post Score	Post-Learning Gain
RMSE	1.586	0.682
Correlation	0.322	0.294
Selected Features	- 2.087 * numLevelsSkippedByChoice - 0.008 * numWrongLevel11 - 0.028 * numWrongLevel12	0.414 * avgLevelAvgRT + 0.065 * numLevelsSkippedByPush - 0.001 * numCorrectLevel3 + 0.004 * numCorrectLevel4 + 0.493 * avgReactionTimeLevel2 - 0.428 * avgReactionTimeLevel13

Table 2 (continued). Summary of the predictive models

	Delayed Post Score	**Delayed Post-Learning Gain**
RMSE	1.540	0.470
Correlation	0.303	0.260
Selected Features	2.189 * avgLevelAvgRT - 0.268 * numLevelsSkippedByPush - 0.012 * numWrongLevel3 + 0.006 * numCorrectLevel12 - 2.154 * avgReactionTimeLevel3 + 1.693 * stdReactionTimeLevel3 - 2.306 * stdReactionTimeLevel12 - 0.528 * avgReactionTimeLevel12	0.841 * avgLevelAvgRT - 0.262 * highestLevelAvgRT + 0.001 * totalWrongHits + 1.434 * avgCorrectLevel1 - 0.003 * numWrongLevel3 + 0.007 * numCorrectLevel12 - 0.009 * numWrongLevel12 - 0.611 * stdReactionTimeLevel12

The models that only used game-level features had a low performance. Excluding level-specific features only did not greatly affect model goodness when predicting the post score, with a correlation of 0.308 and RMSE of 1.589. Features included *numLevelsSkippedByChoice*, *totalWrongHits*, and *numLevelsPlayed*.

6. DISCUSSION AND CONCLUSIONS

Playing AYCET significantly improved students' EF. The effect sizes of EF gains were medium, and that for the delayed post-test 4-6 weeks later was larger than that for the post-test. This difference in effect sizes may be attributed to either the long-term intervention effect by the EF game or students' natural development of EF. Though more evidence is needed, long-term effects of cognitive skills training have been found [18, 32].

We explored the possibility of a game-based assessment of EF using a game originally designed to train EF. We present four linear regression models that use the log data to predict students' EF score of shifting in the post-test, delayed post-test, and the relevant learning gain scores. With correlations around 0.3, these models achieved good performance for preliminary work. This corresponds to the second challenge of this study, which is to separate log data related to EF training from log data related to play in the game context. Good performance of predictive models indicates that a learning game is a promising tool to measure EF.

We generated an extensive list of game-level and level-specific features consisting of accuracy and reaction time indicators. Both accuracy and reaction time features are important in predicting EF but are two potentially distinct dimensions of EF. Generally, at the game level, a moderately higher reaction time and a more consistent reaction time (while controlling for other factors) are positively associated with EF. In addition, a lower reaction time and perhaps a more consistent reaction time are positively associated with EF. As for accuracy features, both correct hits and wrong hits are important for predicting EF.

In addition to accuracy and reaction time features, the number of levels skipped, particularly by the player, was indicative of EF. This means that player progression and player performance are both important for predicting EF.

Most selected features are from level 3 and level 12. This may suggest the key time window, which is the moment after students become familiar with the game mechanics and the moment after a drastic change in difficulty (recall the mandatory push; see section 3.3) may best demonstrate their ability to perform shifting. Varying the difficulty of levels or allowing for some time for students to achieve level 12 may contribute to a better game-based assessment of EF.

Responding to the first challenge, namely, the granularity and time scale for prediction, we found that level-specific features provide more promising results than game-level features only. It is worth further exploring variables at the action-level.

7. IMPLICATIONS AND FUTURE WORK

We explored the techniques of feature engineering and model training to investigate the game-based assessment of EF. The model performance is promising among studies that relate log data with a post-test measure in learning games [33, 34]. Another implication of this work is it sets the foundation for the real-time detection of EF and may provide the basis for dynamic interventions.

Limitations in the current work inspire us to explore more possibilities of game-based assessments for EF. First, a level may be played multiple times by a student. In the current study, all attempts of the same level were aggregated. In the future, we will distinguish multiple attempts of the same level by generating features such as the number of attempts and performance change over attempts. Second, we found that students' performance one or two levels after a challenging level is important. This game mechanic of difficulty change may not apply to other games. We have tried to interpret the model in the context of the specific game design. Third, students experienced a mandatory push in this study (see section 3.3). This is perhaps why features for level 12 were selected. To examine the generalizability of our findings, we will compare the prediction models built under two conditions, one of which replicates the push, while one does not; however, for practical reasons, it is also of interest to determine whether features that only cover earlier levels can predict the post-test and delayed post-test scores as this would require less game play for the assessment. Fourth, we filled the missing data with the average value of a feature. We did so by assuming data were missing at random. More robust methods, such as multiple imputations, could be used moving forward.

In the future, we are interested in generating theory-based features at the action-level (i.e., alien-level) per student, hoping to allow for the real-time detection of EF and for an even better model. An action-level feature may be a student's change in accuracy and reaction time within the first three aliens when the rule changes within a level. Another action-level feature may be the performance curve under different rules within a level. Both features align with the definition of the EF sub-skill of shifting and are not tied to specific levels, so they may produce more generalizable results. Methodologically, we are also interested in comparing the linear regression with other models, such as Support Vector Machines or the Random Forest. Substantively, it may be worth considering accuracy and reaction time as separate outcomes given research suggesting each contribute uniquely to performance on EF tasks in young children [9]. Further work may also apply methods of student modeling to other EF sub-skills, such as inhibition and updating in games that target these skills.

8. ACKNOWLEDGMENTS

Special thanks to Dr. Ryan Baker and Dr. Luc Paquette for their constructive suggestions for the data analysis. This work was supported in part by a research grant from the Institute of Educational Sciences, US Department of Education (Grant #R305A150417). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We would like to thank the students and teachers who participated in this research.

9. REFERENCES

[1] Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., and Fernández-Manjón, B. 2020. Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *Journal of Computer Assisted Learning*. 36, 3, 350-358.

[2] Anguera, J. A. and Gazzaley, A. 2015. Video games, cognitive exercises, and the enhancement of cognitive abilities. *Current Opinion in Behavioral Sciences*. 4, 160-165.

[3] Banich, M. T. 2009. Executive function: The search for an integrated account. *Current Directions in Psychological Science*. 18, 2, 89-94.

[4] Best, J. R. 2012. Exergaming immediately enhances children's executive function. *Developmental Psychology*. 28, 5, 1501-1510.

[5] Best, J. R. and Miller, P. H. 2010. A developmental perspective on executive function. *Child Development*. 81, 6, 1641-1660.

[6] Best, J. R., Nagamatsu, L. S., and Liu-Ambrose, T. 2014. Improvements to executive function during exercise training predict maintenance of physical activity over the following year. *Frontiers in Human Neuroscience*. 8, 353.

[7] Bryce, D., Whitebread, D., and Szűcs, D. 2015. The relationships among executive functions, metacognitive skills and educational achievement in 5 and 7 year-old children. *Metacognition and Learning*. 10, 2, 181-198.

[8] Camerota, M., Willoughby, M. T., Magnus, B. E., and Blair, C. B. 2020. Leveraging item accuracy and reaction time to improve measurement of child executive function ability. *Psychological Assessment*. 32, 12, 1118-1132.

[9] Cardoso-Leite, P., Joessel, A., and Bavelier, D. 2020. 18 Games for enhancing cognitive abilities. In Plass, J. L., Mayer, R. E., and Homer, B. D. ed. *Handbook of Game-Based Learning*. MIT Press, 437-468.

[10] Fuhs, M. W., Nesbitt, K. T., Farran, D. C., and Dong, N. 2014. Longitudinal associations between executive functioning and academic skills across content areas. *Developmental Psychology*. 50, 6, 1698-1709.

[11] Greiff, S., Wüstenberg, S., and Avvisati, F. 2015. Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers and Education*. 91, 92-105.

[12] Hagler, S., Jimison, H. B., and Pavel, M. 2014. Assessing executive function using a computer game: Computational modeling of cognitive processes. *IEEE Journal of Biomedical and Health Informatics*. 18, 4, 1442-1452.

[13] Hake, R. R. 1998. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*. 66, 1, 64-74.

[14] Henderson, N., Kumaran, V., Min, W., Mott, B., Wu, Z., Boulden, D., ... and Lester, J. 2020. Enhancing student competency models for game-based learning with a hybrid stealth assessment framework. In *Proceedings of the 13th International Conference on Educational Data Mining*. International Educational Data Mining Society, 92-103.

[15] Homer, B. D. and Plass, J. L. 2014. Level of interactivity and executive functions as predictors of learning in computer-based chemistry simulations. *Computers in Human Behavior*. 36, 365-375.

[16] Homer, B. D., Plass, J. L., Rafaele, C., Ober, T. M., and Ali, A. 2018. Improving high school students' executive functions through digital game play. *Computers and Education*. 117, 50-58.

[17] Homer, B. D., Plass, J. L., Rose, M. C., MacNamara, A. P., Pawar, S., and Ober, T. M. 2019. Activating adolescents' "hot" executive functions in a digital game to train cognitive skills: The effects of age and prior abilities. *Cognitive Development*. 49, 20–32.

[18] Jaeggi, S. M., Buschkuehl, M., Jonides, J., and Shah, P. 2011. Short-and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*. 108(25), 10081-10086.

[19] Karumbaiah, S., Baker, R. S., & Shute, V. 2018. Predicting quitting in students playing a learning game. In *Proceedings of the 11th International Conference on Educational Data Mining*. International Educational Data Mining Society, 167-176.

[20] Ke, F. and Shute, V. 2015. Design of game-based stealth assessment and learning support. *Serious Games Analytics*. (2015), 301-318.

[21] Khenissi, M. A., Essalmi, F., Jemni, M., Chang, T. W., and Chen, N. S. 2016. Unobtrusive monitoring of learners' interactions with educational games for measuring their working memory capacity. *British Journal of Educational Technology*. 48, 2, 224-245.

[22] Mayer, R. E., Parong, J., and Bainbridge, K. 2019. Young adults learning executive function skills by playing focused video games. *Cognitive Development*. 49, 43-50.

[23] Mislevy, R. J., Steinberg, L. S., and Almond, R. G. 2003. focus article: on the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives. 1, 1, 3-62.

[24] Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. 2000. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive Psychology. 41, 1, 49-100.

[25] Ober, T. M., Brenner, C. J., Olsen, A., Homer, B. D., and Plass, J. L. 2021. Detecting patterns of engagement in a digital cognitive skills training game. *Computers and Education*. 165, 104144.

[26] Owen, V. E., Roy, M. H., Thai, K. P., Burnett, V., Jacobs, D., Keylor, E., and Baker, R. S. 2019. Detecting wheel-spinning and productive persistence in educational games. In *Proceedings of the 12th International Conference on Educational Data Mining*. International Educational Data Mining Society, 378-383.

[27] Parong, J., Mayer, R. E., Fiorella, L., MacNamara, A., Homer, B. D., and Plass, J. L. 2017. Learning executive function skills by playing focused video games. *Contemporary Educational Psychology*. 51, 141-151.

[28] Plass, J. L., Homer, B. D., and Kinzer, C. K. 2015. Foundations of game-based learning. *Educational Psychologist*. 50, 4, 258-283.

[29] Plass, J. L., Homer, B. D., Pawar, S., Brenner, C., and MacNamara, A. P. 2019. The effect of adaptive difficulty adjustment on the effectiveness of a game to develop executive function skills for learners of different ages. *Cognitive Development*. 49, 56-67.

[30] Rowe, E., Almeda, M. V., Asbell-Clarke, J., Scruggs, R., Baker, R., Bardar, E., and Gasca, S. 2021. Assessing implicit computational thinking in zoombinis puzzle gameplay. *Computers in Human Behavior*. 120, 106707.

[31] Rowe, E., Asbell-Clarke, J., Baker, R. S., Eagle, M., Hicks, A. G., Barnes, T. M., ... and Edwards, T. 2017. Assessing implicit science learning in digital games. *Computers in Human Behavior*. 76, 617-630.

[32] Schwaighofer, M., Fischer, F., and Bühner, M. 2015. Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist*. 50, 2, 138–166.

[33] Shute, V. J. and Rahimi, S. 2021. Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*. 116, 106647.

[34] Shute, V. J., Wang, L., Greiff, S., Zhao, W., and Moore, G. 2016. Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*. 63, 106-117.

[35] Valladares-Rodríguez, S., Pérez-Rodríguez, R., Anido-Rifón, L., and Fernández-Iglesias, M. 2016. Trends on the application of serious games to neuropsychological evaluation: A scoping review. *Journal of Biomedical Informatics*. 64, 296-319.

[36] Zelazo, P. D. 2006. The dimensional change card sort (DCCS): A method of assessing executive function in children. *Nature Protocols*. 1, 1, 297-301.

[37] Zelazo, P. D., and Bauer, P. J. 2013. ed. *National Institutes of Health Toolbox Cognition Battery (NIH Toolbox CB): Validation for Children Between 3 and 15 Years*. Wiley, Hoboken, NJ.