Analysing Game Data to Evaluate Children's Learning Progress with the EDURINO App

Ning Ding[1 2], Katherine Chen[2], Max Van Radecke[2], Irene Klemm[2], Franziska Meyer[2] and Nicky Clayton [1]

[1]Department of Psychology, University of Cambridge, Cambridge, UK
[2]EDURINO GmbH, Munich, Germany

ning@edurino.com/nd422@cam.ac.uk
kit@edurino.com
max@edurino.com
irene@edurino.com
franziska@edurino.com
nsc22@cam.ac.uk

Abstract: Game data offers a unique way to examine children's learning progress in a gameplay context. Despite its relevance in evaluating the efficacy of game-based learning, this area is less researched than engagement and usability study. This paper presents a quantitative investigation of EDURINO, an educational gaming app targeting children aged 4-8. We implemented learning metrics to track children's performance in EDURINO games. Specifically, the measurements included standardised variables from educational and experimental psychology, such as error count and duration. Here we reported data from a total of 18 EDURINO minigames in Literacy and Maths. Game data was collected by the learning metrics tracking system. A total of 66,463 players' data was analysed using generalised linear mixed models with the *lm4* package in R. Overall, children have shown improvements while playing the EDURINO app. We examined the relationships between children's game performance, player demographics and game-specific features and discovered different patterns. First, the significant main effects of age and gender indicated that EDURINO's educational content was developmentally sensitive and age appropriate. Second, children made fewer errors and spent less time when they were in the later rounds within a game or have played the same game more than once. Third, higher difficulty level undermined game performance. Worth noting is that the effect of game-specific factors varied among different educational content, and such discrepancy could be attributed to the nature of game design. When presented with engaging and interactive content, children may have chosen to deliberately make mistakes for fun effects. These findings are discussed in relation to game-based assessment for learning in EdTech research and industry. We also considered the interpretation of learning progress in EDURINO's gameplay context while highlighting the need for further efficacy research.

Keywords: Game-based Learning, Educational Technology, Game Data, Learning Progress

1. Introduction

The rapid integration of technology and digital devices into children's life has fundamentally transformed educational practices. With the EdTech industry estimated to reach $342 billion by 2025 (IQ Holon), there is an increasing emphasis on the development of digital platforms to enhance learning outcomes. The combination between engaging educational content and evidence-based pedagogical principles makes the field of game-based learning particularly promising (Behnamnia et al., 2022; Hirsh-Pasek et al., 2015; Tang, 2020). Indeed, cumulative research has demonstrated significant learning benefits of educational apps, with a meta-analysis revealing a medium positive effect size of +0.31, particularly in enhancing young children's mathematics and literacy skills (Griffith et al., 2020; Kim et al., 2021).

However, a critical gap remains in understanding the mechanism of specific educational apps. Traditional efficacy research predominantly employed methods of training study, survey, and randomised-controlled trials. However, few studies have conducted in-depth analysis of the educational tools themselves, specifically, to show how real-life player data can be utilised to assess and measure learning within the gameplay context (Bang, Li and Flynn, 2023; Fisch et al., 2024; Walsh and Bokhove, 2021). The current study aims to fill this gap by investigating real-life player data from the EDURINO's app.

EDURINO is an educational gaming app for children aged 4-8. There are nine subject areas covering STEM, humanities, and social-emotional learning in both German and English. Its unique features include a 'magic pen' stylus and haptic figurines that bridge digital and real-world learning (Figure 1). EDURINO's educational content has both holistic as well as detailed scope and sequences of learning objectives and the structure is informed by pedagogical theories (Plass and Pawar, 2020). In EDURINO's competency-based learning context, children are given small activities to learn and practice specific knowledge. Also, all materials are presented in a playful and engaging digital platform with clear context and narratives.

To examine the gameplay data from EDURINO, this study was informed by different methodological frameworks of learning processes within digital games (Baker, Gašević, and Karumbaiah, 2021; Siemens and Baker, 2012). Notably, the entitative/reductionist paradigm has been the mostly widely adopted in research on educational apps (Fisch et al., 2024; Halverson and Owen, 2014). It focuses on breaking down complex phenomena, i.e., learning within a game, into their constituent components, e.g., game design and player behaviour, to understand relationships between these factors. Nevertheless, the uncertainty and noise in game-based assessment data require detailed examination of the game content and mechanics to establish reliable and meaningful learning metrics (Xu et al., 2021).

Figure 1: EDURINO's haptic figurines (upper left), game played with the "magic pen" on a tablet (upper right), screenshots of a literacy game (lower left), and a maths game (lower right).

The current study is the first empirical research utilising EDURINO's gameplay data. Firstly, we adopted the reductionist approach to define learning metrics while accounting for specific game design and mechanics. Second, we internally implemented the tracking system in the EDURINO app to collect data from real players. Given the exploratory nature of this study, we did not make specific and prior assumptions.

Building on the identified gaps and challenges, the current study explored two research questions:

1. Is there any evidence to suggest that children are learning within EDURINO's gameplay context?
2. What are the relationships between player features (demographics), game-specific features (design and content), and children's game performance in the EDURINO app?

2. Material and Methods

2.1 Participants and Ethics

Participants were active registered users of the EDURINO app (DACH version). The collection and report of children's game data did not violate the data privacy statement, since it was used only for research purpose, and all analyses were conducted by EDURINO's internal researcher with no third-party involvement. Furthermore, all procedures performed in the current study were in accordance with the ethical standards of the University of Cambridge Psychology Research Ethics Committee and have been reviewed by the senior author as an independent consultant on the project.

2.2 Data Source

The study utilised real-life player data of children's game performance, which was collected by the learning metrics tracking system and stored in a centralised cloud database. In EDURINO's learning world, there are 12 different episodes/subjects and each episode consists of numerous minigames which cover multiple learning objectives of that subject. Within each minigame, there are at least two rounds of activities/tasks that children need to complete. Notably, the number of rounds for each minigame is fixed by design. This paper focused on a subset of EDURINO games, specifically 10 minigames from Episode 1 Mika Literacy in German (E1) and 8 minigames from Episode 2 Robin Mathematics (E2). The learning objectives of each minigame were outlined in Supplementary Material (Table S1). Data reported in this study was collected from 29th May 2024 to 1st July 2024. Data collection was conducted remotely through the implemented tracking system, in that player data was collected automatically while children were playing the app. There was no fixed duration for the study since children interacted with the app at their own pace and choice.

2.3 Measures

Multiple measures were first defined then tracked for each minigame (Table 1). The unit of collection is by round. Each round/minigame is designed to teach and test children's knowledge of specific learning objectives and they are given a task/activity to complete. In addition, we extracted information of age and gender for each player, which was entered by parents when setting up a child's player profile.

Table 1: Definition of measures.

Measure	Definition
Error	The total number of errors the child makes in a round, e.g., choose the wrong symbol/quantity to complete a mathematical equation.
Duration (in seconds)	The total length the child takes to successfully pass a round.
Round number	The specific, discrete and ordered segment of a minigame; the child needs to successfully complete certain task/activities in a round in order to move to the next round.
Difficulty level	The level of how difficult the learning content is in a round, set by curriculum and game designers. The higher the difficulty level, the more challenging the content.
Play count	The number of how many times the child has played a particular minigame, play count = 1 indicates first-play and so on.

2.4 Analytic Plan

To investigate how changes of children's performance in EDURINO games were associated with demographic variables and game-related factors, we fitted a series of Generalised Linear Mixed Models (GLMM) with the lm4 packages in R (Bates et al., 2015). For each response variable, we ran two models. Specifically, the first model only examined children's first play experience (play count = 1) while the second model utilised the full dataset (all play count included). In E1, we fitted models with fixed effects of age group (6 groups, 3-8 years), gender (male, female), round number, and play count (only for model 2), while participant ID was entered as a random effect. Based on market research, we extended the age range to include 3-year-olds to account for this customer group in the analysis. Notably, for E2, we included *difficulty level*, rather than round number, as fixed effects in the models. Minigames in E1 do not have varying difficulty levels whereas in E2 there are adaptable difficulty levels based on children's game performance. Taken the multicollinearity problem into account, we decided against fitting the models using both round number and difficulty level simultaneously. Therefore, the final fixed effects for all E2 minigames were age group (6 groups, 3-8 years), gender (male, female), difficulty level, and play count (only for model 2), while participant ID was entered as a random effect. Moreover, no interaction terms were included given the exploratory nature of the current study and the scarcity of relevant literature. Steps on data preparation are explained in Supplementary Material (https://figshare.com/articles/online_resource/Supplementary_Material_for_the_full_length_paper/26395765?file=48349339).

3. Results

Descriptive statistics of learning metrics measures in E1 and E2 were outlined in Table 2. The breakdown of player demographics was presented in Supplementary Material (Table S2). In GLMM, a significant estimate of fixed effect indicates that a particular independent variable, i.e., a game-specific factor, has a statistical meaningful impact on the outcome variable, i.e., error count and duration per round.

3.1 Children's Game Performance in E1

In children's first play, age group, gender, and round number were all significant main effects for predicting children's game performance (Table 3, 4). Older children made fewer errors and spent shorter duration to pass a round than the younger group. Also, children made fewer errors and spent less time with the later rounds than the earlier rounds within the same minigame, except for Boss Fight 1 & 2, and Rhyme Wolverine. Moreover, female outperformed male in all minigames, except for Boss Fight 6 and Bridge of Word Holes.

Table 2: Descriptive statistics of learning metrics measures. *N* refers to the number of players.

Minigame	Round number	Difficulty level	Mean of error	Mean of duration
Episode 1 Mika Literacy				
Boss Fight 1 (*N* = 5869)	6	-	0.485	18.601
Boss Fight 2 (*N* = 4398)	5	-	0.194	21.482
Boss Fight 6 (*N* = 2284)	5	-	0.320	11.691
Bridge with Word Holes (*N* = 5427)	8	-	0.129	13.312
Burping Fish (*N* = 4098)	3	-	0.736	41.543
In the Head of Ava (*N* = 5414)	8	-	0.122	11.027
Rhyme Wolverine (*N* = 8077)	2	-	2.200	62.849
Wheel of Sounds 1* (*N* = 2957)	8	-	1.011	35.733
Wheel of Sounds 2* (*N* = 2308)	8	-	1.230	32.351
Wheel of Sounds 3* (*N* = 2118)	8	-	1.305	32.739
Episode 2 Robin Mathematics				
Bridge of Numbers (*N* = 2844)	4	2	0.427	35.227
Capture Shapes (*N* = 4270)	8	4	0.859	19.679
Construction Scale (*N* = 2486)	5	3	0.418	36.775
Put in Numbers (*N* = 3671)	4	2	0.359	30.346
Transport Building Material 1 (*N* = 3737)	5	3	0.206	42.124
Transport Building Material 2 (*N* = 2400)	5	3	0.229	42.749
Treasure Map (*N* = 2053)	6	3	0.095	27.557
Wheel of Fortune (*N* = 2052)	5	3	0.589	38.679

Note. There are three variants of the minigame Wheel of Sounds; each variant has the same learning objectives and game design but was separately tracked.

When all play counts were included, age group, gender, and round number were significant main effects in predicting the number of errors and duration spent in a round (Table 5, 6). Yet, the effect of play count differed among minigames. In Boss Fight 1 & 2, and Rhyme Wolverine, the number of errors increased with higher play count. Whereas for In the Head of Ava, children made less errors when they have played the minigame more than once. Similarly, higher play count was significantly related to shorter duration in all the minigames except for Boss Fight 6. To visualise the relationship between each main effect predictor and response variables, we generated graphs based on the GLMM models using the *ggeffects* package in R (Lüdecke, 2018). For instance, Figure 2 illustrates the marginal effect of how each predictor and duration was associated in Boss Fight 1. All figures for E1 minigames were presented in *Figshare* (*https://figshare.com/articles/figure/E1_marginal_effect_figures/26349118*).

Table 3: GLMM models (error as response variable and play count is 1) in E1.

	Estimates of fixed effects for error			
Minigame	Age group	Gender	Round number	Constant
Boss Fight 1	-0.294***	-0.169***	0.093***	-0.203**
Boss Fight 2	-0.123***	-0.039	0.054***	-1.520***
Boss Fight 6	-0.402***	-0.064	-0.093***	0.948***
Bridge with Word Holes	-0.350***	-0.068	-0.048***	-0.715***
Burping Fish	-0.208***	-0.082**	-0.050***	0.724***
In the Head of Ava	-0.240***	-0.088**	-0.075***	-0.940***
Rhyme Wolverine	-0.237***	-0.146***	0.086***	1.001***
Wheel of Sounds 1	-0.246***	-0.114**	-0.163***	1.652***
Wheel of Sounds 2	-0.243***	-0.091**	-0.219***	2.078***
Wheel of Sounds 3	-0.255***	-0.079**	-0.224***	2.231***

Note. *** - p < .001** - p < .01* - p < .05.

Table 4: GLMM models (duration as response variable and play count is 1) in E1.

	Estimates of fixed effects for duration			
Minigame	Age group	Gender	Round number	Constant
Boss Fight 1	-0.085***	-0.007	-0.022***	3.453***
Boss Fight 2	-0.085***	-0.021*	-0.076***	3.747***
Boss Fight 6	-0.103***	0.003	0.032***	2.870***
Bridge with Word Holes	-0.067***	0.002	-0.018***	3.015***
Burping Fish	-0.056***	-0.024***	-0.044***	4.135***
In the Head of Ava	-0.059***	-0.006	-0.053***	2.987***
Rhyme Wolverine	-0.065***	-0.016*	-0.167***	4.800***
Wheel of Sounds 1	-0.074***	-0.012	-0.061***	4.244***

	Estimates of fixed effects for duration			
Minigame	**Age group**	**Gender**	**Round number**	**Constant**
Wheel of Sounds 2	-0.061***	0.001	-0.085***	4.151***
Wheel of Sounds 3	-0.055***	0.005	-0.077***	4.099***

Note. *** - p < .001** - p < .01* - p < .05.

Table 5: GLMM models (error as response variable) in E1.

	Estimates of fixed effects for error				
Minigame	**Age group**	**Gender**	**Round number**	**Play count**	**Constant**
Boss Fight 1	-0.280***	-0.146***	-0.002	0.293***	-0.139*
Boss Fight 2	-0.132***	0.016	-0.013	0.190***	-1.422***
Boss Fight 6	-0.401***	-0.068	-0.102***	-0.027	0.999***
Bridge with Word Holes	-0.318***	-0.061	-0.051***	-0.010	-0.825***
Burping Fish	-0.199***	-0.051**	-0.029**	-0.023	0.670***
In the Head of Ava	-0.201***	-0.097**	-0.074***	-0.209***	-0.916***
Rhyme Wolverine	-0.257***	-0.146***	0.070***	0.161***	1.095***
Wheel of Sounds 1	-0.232***	-0.092***	-0.167***	0.016	1.594***
Wheel of Sounds 2	-0.238***	-0.069*	-0.218***	0.003	2.038***
Wheel of Sounds 3	-0.257***	-0.074**	-0.221***	-0.008	2.229***

Note. *** - p < .001** - p < .01* - p < .05.

Table 6: GLMM models (duration as response variable) in E1.

	Estimates of fixed effects for duration				
Minigame	**Age group**	**Gender**	**Round number**	**Play count**	**Constant**
Boss Fight 1	-0.079***	-0.009	-0.014***	-0.095***	3.483***
Boss Fight 2	-0.081***	-0.015	-0.050***	-0.121***	3.769***
Boss Fight 6	-0.103***	-0.004	0.032***	0.001	2.874***
Bridge with Word Holes	-0.058***	-0.005	-0.006***	-0.044***	2.961***
Burping Fish	-0.047***	-0.020***	0.002	-0.051***	4.035***
In the Head of Ava	-0.046***	-0.013***	-0.036***	-0.095***	2.931***
Rhyme Wolverine	-0.055***	-0.021***	-0.085***	-0.104***	4.698***
Wheel of Sounds 1	-0.068***	-0.010**	-0.055***	-0.117***	4.296***
Wheel of Sounds 2	-0.057***	0.002	-0.081***	-0.070***	4.196***
Wheel of Sounds 3	-0.055***	-0.007***	-0.074***	-0.073***	4.156***

Note. *** - p < .001** - p < .01* - p < .05.

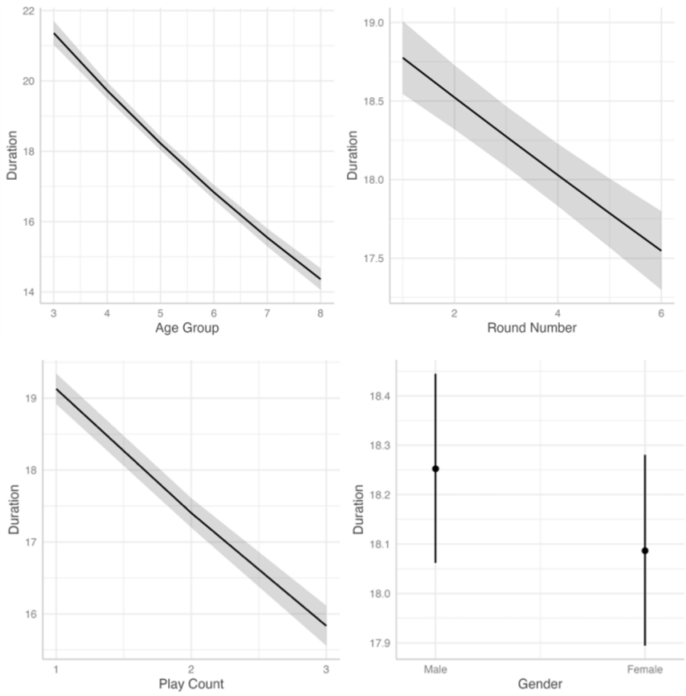

Figure 2: Predicted marginal effects of age group, round number, play count, and gender on duration spent for BOSS Fight 1, with 95% confidence intervals. Each chart illustrates how changes in a specific main effect are associated with variations with duration, holding the other main effects constant.

3.2 Children's Game Performance in E2

Irrespective of play count, age group was a consistent and significant main effect predictor for children's game performance, i.e., error count and duration, across all the minigames (Table 7, 8). Similar to E1, gender was the second most consistent main effect: female made less errors and took less time to pass a round. In terms of difficulty level, when play count = 1, higher difficulty level was associated with more errors or longer duration in Bridge of Numbers, Put in Numbers, Wheel of Fortune, and Capture of Shapes. However, we also found the opposite and less expected pattern: higher difficulty level was associated with fewer errors or shorter duration in Construction Scale, Transport Building Material 1&2, and Treasure Map.

When all play counts were included for analysis, higher difficulty level was associated with fewer error count in Put in Numbers and Transport Building Material 1 (Table 9). Yet, for Bridge of Numbers, Capture Shapes, and Wheel of Fortune, higher difficulty level was associated with poorer game performance, i.e., greater errors and longer duration to pass a round. Higher play count was a significant predictor for shorter duration for all minigames except for Wheel of Fortune. However, the pattern was less clear with the error count; higher play count was associated with better performance in Construction Scale, Put in Numbers, Transport Building Material 1, yet related to more errors in Capture Shapes and Transport Building Material 2. We followed the same procedure as in E1 to generate figures to visualise GLMM model results. For instance, Figure 3 illustrates the marginal effect of all fixed effects predictors of error count in Put in Numbers. All E2 figures were presented in *Figshare*

(https://figshare.com/articles/figure/_sup_E2_marginal_effect_figures_sup_/26349130)

Table 7: GLMM models (error as response variable and play count is 1) in E2.

	Estimates of fixed effects for error			
Minigame	**Age group**	**Gender**	**Difficulty level**	**Constant**
Bridge of Numbers	-0.155***	-0.060	0.265***	-0.767***
Capture Shapes	-0.186***	-0.072***	0.014	0.535***
Construction Scale	-0.150***	-0.038	-0.202***	0.231*
Put in Numbers	-0.180***	-0.045	0.021	-0.477***
Transport Building Material 1	-0.232***	-0.25	-0.310***	0.183
Transport Building Material 2	-0.219***	-0.258***	0.014	-0.998***
Treasure Map	-0.225***	-0.216**	0.020	-1.936***
Wheel of Fortune	-0.239***	-0.126**	0.391***	-0.327***

Note. *** - p < .001** - p < .01* - p < .05.

Table 8: GLMM models (duration as response variable and play count is 1) in E2.

	Estimates of fixed effects for duration			
Minigame	**Age group**	**Gender**	**Difficulty level**	**Constant**
Bridge of Numbers	-0.063***	-0.001	0.280***	3.443***
Capture Shapes	-0.108***	-0.046***	0.101***	3.220***
Construction Scale	-0.061***	-0.020	-0.050***	4.025***
Put in Numbers	-0.063***	-0.0001	0.169***	3.396***
Transport Building Material 1	-0.065***	0.009	-0.235***	4.633***
Transport Building Material 2	-0.069***	-0.003	-0.257***	4.699***
Treasure Map	-0.050***	-0.014	-0.112***	3.849***
Wheel of Fortune	-0.079***	-0.018	0.355***	3.366***

Note. *** - p < .001** - p < .01* - p < .05.

Table 9: GLMM models (error as response variable) in E2.

	Estimates of fixed effects for error				
Minigame	**Age group**	**Gender**	**Difficulty level**	**Play count**	**Constant**
Bridge of Numbers	-0.165***	-0.035	0.292***	0.001	-0.753***
Capture Shapes	-0.187***	-0.049*	0.035***	0.029*	0.364***
Construction Scale	-0.159***	-0.0002	-0.048	-0.151***	0.032
Put in Numbers	-0.183***	-0.055	-0.216***	-0.084***	0.077
Transport Building Material 1	-0.230***	0.021	-0.144***	-0.295***	-0.313*
Transport Building Material 2	-0.233***	-0.299***	0.042***	0.027***	-0.958***
Treasure Map	-0.193***	-0.130	0.038	0.034	-2.076***
Wheel of Fortune	-0.229***	-0.070	0.259***	0.037	-0.160***

Note. *** - p < .001** - p < .01* - p < .05.

Table 10: GLMM models (duration as response variable) in E2.

	Estimates of fixed effects for duration				
Minigame	**Age group**	**Gender**	**Difficulty level**	**Play count**	**Constant**
Bridge of Numbers	-0.059***	-0.0002	0.296***	-0.029***	3.427***
Capture Shapes	-0.102***	-0.033***	0.100***	-0.034***	3.223***
Construction Scale	-0.059***	-0.012	0.017***	-0.082***	3.941***
Put in Numbers	-0.056***	0.004	0.011	-0.044***	3.701***
Transport Building Material 1	-0.067***	0.032***	-0.101***	-0.167***	4.478***
Transport Building Material 2	-0.073***	0.001	-0.125***	-0.146***	4.550***
Treasure Map	-0.052***	-0.005	-0.036***	-0.065***	3.735***
Wheel of Fortune	-0.073***	-0.010	0.177***	-0.001	3.678***

Note. *** - p < .001** - p < .01* - p < .05.

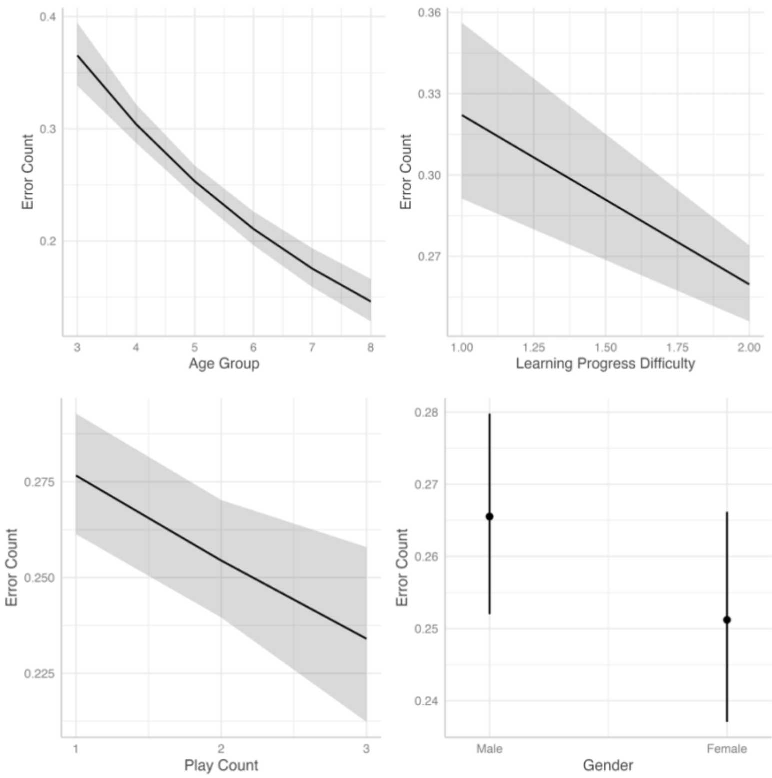

Figure 3: Predicted marginal effects of age group, difficulty level, play count, and gender on error count for Put in Numbers, with 95% confidence intervals. Each plot illustrates how changes in a specific main effect are associated with variations with error count, holding the other main effects constant.

4. Discussion

The current study was the first investigation to utilise EDURINO's game data to examine children's learning progress in the gameplay context. First, based on standardised practice from educational and experimental psychology, we defined an array of measures for learning metrics. Then, we implemented the tracking system to collect data of children's performance in EDURINO's educational games. Overall, children demonstrated learning progress within EDURINO's gameplay context, and their performance was closely related to both player demographics as well as game-specific features.

In answering the first research question, interpretation of game performance between rounds and play count warrant discussion. Across all minigames in Episode 1 for literacy, round number was a significant main effect in GLMM. Specifically, children made fewer errors and spent less time to pass the later rounds than the earlier rounds within the same minigame. The findings indicate improvements within a very short amount of time, and therefore by themselves may not be seen as conclusive evidence of learning. Play count, which has been observed over a longer course of time, was also associated with clear improvements in children's game performance. Children continued the trend of making less mistakes and taking less time to pass when they have played the game more than once. Worth highlighting is that although rounds within the same minigame have the same game mechanics and learning objectives, children are presented with different content in each round, for example detecting different quantities. Therefore, children's improvements could not be explained entirely by repetition or practice effect (Fisch et al., 2024).

Taken together, EDURINO's game data was indicative of learning progress within the gameplay context. Nevertheless, game data are at best, assessment of constrained skills on specific skills such as sorting shapes and counting, rather than an assessment of unconstrained skills of broader domain knowledge such as problem solving. As noted by Kim et al. (2021), educational apps show constrained learning benefits with a small and fixed body of knowledge. To address this gap, more research is needed to test EDURINO's efficacy with transferred contexts and broader materials to show unconstrained learning benefits (Aladé et al., 2016; Pearson et al., 2020).

To address the second research question, we fitted several main effects in GLMM models to elucidate the relationships between player and game-specific factors on children's game performance. Not only did the findings provide insights into EDURINO's game design, but also highlighted the challenges for game-based learning and assessment. First, across all minigames assessed in the study, older children consistently outperformed younger children. Additionally, there were also clear gender differences in that female exhibited better game performance than male, particularly in E1 for literacy games. The significant main effects of age and gender suggested that EDURINO's learning content and educational materials were developmentally sensitive and age appropriate, since it follows the universal age and gender-related patterns and trajectories (Eriksson et al., 2012; Bang, Li and Flynn, 2023).

Difficulty level, on the other hand, led to more mixed results. In some minigames, such as Capture Shapes, Construction Scale, Treasure Map, and Transport Building Material, children made fewer mistakes and spent shorter duration with content of higher difficulty level. There were also cases where children started to make more errors yet spent a shorter time when they have played the game more than once. Several explanations are possible. First, certain games' difficulty level might not correspond to children's ability and skills and need adjustment. Second, children might have already mastered the skills required for the games but started to play for fun by deliberately making mistakes. EDURINO incorporates the 'giggle factor' feature in design so that errors are presented and responded with fun and interactive effects. This posits a wider question and challenge of how to reliably assess game-based data for educational content when the content is presented in a gameplay context (Fisch et al., 2024; Walsh and Bokhove, 2021). Using data mining techniques, advanced tracking system and analytic approach would be powerful tools to account for the noise in game data (Baker, Corbett, and Aleven, 2008; Chen et al., 2020).

A particularly novel contribution of the current study was utilising game data from real players to examine game-based learning, a largely under researched field with very little literature (Behnamnia et al., 2022; Kiguchi, Saeed, and Medi, 2022; Kucirkova et al., 2014; Walsh and Bokhove, 2021). The clear and precise definition of more than one outcome measures (of error count and duration) allows for a richer interpretation of findings and could be adopted by other apps in the industry to establish comparable learning metrics (Siemens and Baker, 2012). The very large sample size and robust statistical analysis ensured the results are consistent with academic standards and practice, and we advocate that other commercial companies adopt inferential methods for research design. However, the study relied on game data which is naturally prone to errors and uncertainty (Xu et al., 2021). Most importantly, the findings indicate improvements within EDURINO's gameplay context, further investigation outside the gameplay context, for example longitudinal efficacy research, is essential before any conclusion is drawn about its long-term learning benefits for unconstrained knowledge.

5. Conclusion

The current research focused on game-assessment data to reveal that children shown clear improvements in EDURINO's gameplay context. Learning progress was evident within a single play session as well as across multiple play sessions. The app's engaging content was age appropriate and children's game performance followed universal developmental trajectory. We consider game-based learning as an important piece of the puzzle in understanding the holistic benefits of educational apps. Yet, the significance of traditional efficacy research cannot be underestimated.

References

Aladé, F., Lauricella, A.R., Beaudoin-Ryan, L. and Wartella, E., 2016. Measuring with Murray: Touchscreen technology and preschoolers' STEM learning. *Computers in human behavior*, *62*, pp.433-441.

Baker, R.S., Corbett, A.T. and Aleven, V., 2008. Improving contextual models of guessing and slipping with a truncated training set.

Baker, R.S., Gašević, D. and Karumbaiah, S., 2021. Four paradigms in learning analytics: Why paradigm convergence matters. *Computers and Education: Artificial Intelligence*, *2*, p.100021.

Bang, H.J., Li, L. and Flynn, K., 2023. Efficacy of an adaptive game-based math learning app to support personalized learning and improve early elementary school students' learning. *Early Childhood Education Journal*, *51*(4), pp.717-732.

Bates, D., Kliegl, R., Vasishth, S. and Baayen, H., 2015. Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Behnamnia, N., Kamsin, A., Ismail, M. A. B., & Hayati, S. A. (2023). A review of using digital game-based learning for preschoolers. *Journal of Computers in Education*, *10*(4), 603-636.

Chen, X., Zou, D., Cheng, G. and Xie, H., 2020. Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. *Computers & Education*, *151*, p.103855.

Eriksson, M., Marschik, P.B., Tulviste, T., Almgren, M., Pérez Pereira, M., Wehberg, S., Marjanovič-Umek, L., Gayraud, F., Kovacevic, M. and Gallego, C., 2012. Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British journal of developmental psychology*, *30*(2), pp.326-343.

Fisch, S.M., Fletcher, K., Abdurokhmonova, G., Davis, L., Fisch, N., Fisch, S.R., Jurist, M., Kestin, R., Pesch, A., Seguì, I. and Shulman, J., 2024. "I wonder, what if, let's try": Sesame Street's playful learning curriculum impacts children's problem solving. *Journal of Children and Media*, *18*(3), pp.334-350.

Griffith, S.F., Hagan, M.B., Heymann, P., Heflin, B.H. and Bagner, D.M., 2020. Apps as learning tools: a systematic review. *Pediatrics*, *145*(1).

Halverson, R. and Owen, V.E., 2014. Game-based assessment: An integrated model for capturing evidence of learning in play. *International Journal of Learning Technology*, *9*(2), pp.111-138.

Hirsh-Pasek, K., Zosh, J.M., Golinkoff, R.M., Gray, J.H., Robb, M.B. and Kaufman, J., 2015. Putting education in "educational" apps: Lessons from the science of learning. *Psychological science in the public interest*, *16*(1), pp.3-34.

Holon, I.Q., 10. charts that explain the Global Education Technology Market. *Everything you need to know about the Global Education Technology Market in*, *10*.

Kiguchi, M., Saeed, W. and Medi, I., 2022. Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest. *Applied Soft Computing*, *118*, p.108491.

Kim, J., Gilbert, J., Yu, Q. and Gale, C., 2021. Measures matter: A meta-analysis of the effects of educational apps on preschool to grade 3 children's literacy and math skills. *Aera Open*, *7*, p.23328584211004183.

Kucirkova, N., Messer, D., Sheehy, K. and Panadero, C.F., 2014. Children's engagement with educational iPad apps: Insights from a Spanish classroom. *Computers & Education*, *71*, pp.175-184.

Lüdecke, D., 2018. ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), p.772.

Pearson, P.D., Palincsar, A.S., Biancarosa, G. and Berman, A.I., 2020. Reaping the Rewards of the Reading for Understanding Initiative. *National Academy of Education*.

Plass, J.L. and Pawar, S., 2020. Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, *52*(3), pp.275-300.

Siemens, G. and Baker, R.S.D., 2012, April. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254).

Tang, J.T., 2023. Comparative study of game-based learning on preschoolers' English vocabulary acquisition in Taiwan. *Interactive Learning Environments*, *31*(4), pp.1958-1973.

Walsh, C. and Bokhove, C., 2021. Targeting data collection in games based assessment. *Computers and Education Open*, *2*, p.100054.

Xu, W., Liang, H.N., Yu, K. and Baghaei, N., 2021, May. Effect of gameplay uncertainty, display type, and age on virtual reality exergames. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).