Application of cluster analysis to identify different reader groups through their engagement with a digital reading supplement

Yawen Ma[a]

Kate Cain[b]

Anastasia Ushakova[c]

[a] corresponding author, Centre for Health Informatics, Computing, and Statistics, Lancaster Medical School, Lancaster University, Lancaster, United Kingdom. y.ma24@lancaster.ac.uk

[b] Department of Psychology, Lancaster University, Lancaster, United Kingdom. k.cain@lancaster.ac.uk

[c] Centre for Health Informatics, Computing, and Statistics, Lancaster Medical School, Lancaster University, Lancaster, United Kingdom. a.ushakova@lancaster.ac.uk

Acknowledgements: The authors acknowledge the support of the Economic and Physical Research Council who funded the first author through a PhD studentship, and Boost Amplify for supplying the dataset for analysis.

Declarations of interest: none. We confirm that these are original data, not published nor under review elsewhere.

ACCEPTED FOR PUBLICATION: Computers and Education 24/02/2024

Abstract

The focus of this study is the identification of reader profiles that differ in performance and progression in an educational literacy app. A total of 19,830 students in Grade 2 from 347 Elementary schools located in 30 different districts in the United States played the app from 2020 to 2021. Our aim was to identify unique groups of readers using an unsupervised statistical learning technique - cluster analysis. Six indicators generated from the students' log files were included to provide insights into engagement and learning across four different reading-related skills: phonological awareness, early decoding, vocabulary, and comprehension processes. A key aim was to evaluate the implementation and performance of Gaussian mixture models, k-means, k-medoids, clustering large applications and hierarchical clustering, alongside provision of detailed guidance that can benefit researchers in the field. K-means algorithm performed the best and identified nine groups of readers. Children with low initial reading ability showed greater engagement with code-related games (phonological awareness, early decoding) and took longer to master these games, whereas children with higher initial ability showed more engagement with meaning-related games (vocabulary, comprehension processes). Our findings can inform further research that aims to understand individual differences in learning behaviour within digital environments both over time and across various cohorts of children.

Keywords: Early years education, Data science applications in education, Games.

1 Introduction

National efforts in the U.S. (e.g., Common Core (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010); the No Child Left Behind (The No Child Left Behind Act of 2001, 2002)) have not resulted in substantial improvements in literacy performance (Guterres, 2020). Lower academic performance continues to persist among minority students and those from low-income backgrounds (de Brey et al., 2019; Department for Education, 2019). Using digital tools for reading supplements has been found to result in better outcomes, effectively enhancing standard classroom practices (Biancarosa & Griffiths, 2012). It is therefore critical to understand under what conditions digital learning can support literacy development.

Digital games that aid literacy acquisition have some of the characteristics of computer games, so insights from the analysis of computer game player behaviour data may be informative. First, computer games include progression through different skill levels. Second, they generate a high volume of data, including both accuracy and time data at the item level for each player during engagement with the game, across sessions on multiple days. Tracking user behaviour in real-time, as they play, helps game developers identify popular game elements or design flaws and different players' profiles, all of which can inform product development (Isbister & Schaffer, 2008; Sánchez et al., 2012). The analytic techniques used to study engagement with games designed for entertainment could be applied to educational games, with the results used to drive theory and recommendations for how to use digital supplements beneficially, in the classroom. The aim of this study is to explore reading engagement profiles and their relation to reading outcomes in a digital app.

1.1 Reading and literacy development

Reading comprehension is considered the product of a reader's decoding (word reading) skill and listening comprehension (Gough & Tunmer, 1986; Hoover & Gough, 1990). There is significant evidence of a strong relationship between decoding and reading comprehension, particularly in younger readers (e.g., Language and Reading Research Consortium (LARRC), 2015). In terms of underlying skills, children's progress in decoding is correlated with early phonological awareness (e.g., Adams, 1990; Goswami & Bryant, 1990; National Reading Panel, 2000; Snow et al., 1998), and their progress in listening and reading comprehension is strongly predicted by vocabulary (Anderson & Freebody, 1981) and the comprehension processes skills that support multi-clause sentence and single text comprehension (Oakhill & Cain, 2012). This evidence informed the design of the app's games and provided an empirical and theoretical basis for our exploration of games targeting four skills, namely phonological awareness, early decoding, vocabulary, and comprehension processes. Recent studies have confirmed that digital reading games can aid the development of (some of these) specific skills that underpin reading development, such as phonological awareness and vocabulary (e.g., Hofmann, 2021; López-Escribano et al., 2021; Vnucko & Klimova, 2023) and the need to consider reader characteristics, such as age, ability, gender etc. (e.g., Benton et al., 2023; Diprossimo et al., 2023).

1.2 Data driven potential measures of engagement with educational games

Expert performance does not simply arise from innate aptitude, but from sustained regular deliberate practice, which is often time limited (Ericsson et al., 1993). Regular and active engagement in language learning is encouraged to improve literacy ability (Wyse et al., 2013), and motivation and engagement contribute to improve learning outcomes (Kuh, 2009; Lepper et al., 2005; Schlechty, 2001; Woolfolk, 2007). Students' engagement with a digital learning game can be measured by the behavioural indicators which represent the time and effort they put into it. These can be quantified as the time and frequency of training, number of levels of interaction, etc.

Dynamic behavioural data can accurately describe the time and, by proxy, the effort that students dedicate to a target skill. A relevant study to investigate this dimension of engagement was conducted by Yang and colleagues (2020) who used six interactive e-book user log variables to identify a student's in-the-moment reading processes. They found that log variables, such as time spent reading, frequency of reading the book, and number of attempts to get the correct answer, were associated with word knowledge and strategic reading outcomes. In other work, time spent reading has been found to relate significantly to improvements in pupils' reading achievement (Locher & Pfost, 2020). It has been proposed that time spent reading may indicate 'careful reading' and the depth of involvement in learning word knowledge (Laufer & Hulstijn, 2001). Ciampa (2012) used reading frequency as an indicator of motivation, since students who read more frequently could have higher levels of motivation (Laufer & Hulstijn, 2001). The number of attempts that pupils make to get the right answer reflects their strategic approach; the more attempts they need before they get the right answer, the less strategic they are being, indicating that they are "gaming the system" or guessing and checking (Baker et al., 2004, 2008).

This range of engagement measures has not been explored comprehensively in previous work. Furthermore, how to use and interpret these variables is not clear because they may differ by skill, for example, it is unlikely that there is a single threshold of frequency that is suitable for all skills (Du et al., 2021). No studies, to date, have summarised a generalised set of engagement variables for digital educational games. Furthermore, from the perspective of output variables, none captures information such as correctness or speed of mastery. Therefore, before proceeding to explore the relationship between behaviour and learning outcomes, a suitable way to take into account the effort that students put into the different skills and to rescale the values of engagement variables for different skills on a common scale is needed.

1.3 Relationship between engagement with and performance on educational games

Some studies report a significant correlation between student engagement and learning gains (Agudo-Peregrina et al., 2014; Gómez-Aguilar et al., 2015), and have used student engagement with a specific app to predict learning outcomes, such as the final test results. For example, Zheng et al. (2020) found a positive correlation between the number of logins for online English and literature courses and final outcomes of K-12 students in the U.S. (see also, Liang et al., 2014; Liao & Wu, 2023). This previous work has focused on high-school and undergraduate students, who have already mastered the basics of reading. The current study is unique in its focus on beginner readers, to determine if and how engagement profiles explain performance and progress in the initial stages of reading development.

1.4 Motivation of the study and research questions

Successful reading comprehension relies on many processes including decoding the written words, retrieving their meanings, computing sentence structures, and integrating information across these sentences and with a reader's general knowledge, to build up a coherent representation of the text's meaning (Perfetti et al., 2005). We focus on four skills that underpin two critical aspects of reading development: code-related skills (phonological awareness, early decoding) and meaning-related skills (vocabulary, comprehension processes). We seek to determine if engagement with these two skill sets indicates distinct clusters of children, by utilizing a high-volume and unstructured dataset obtained from an educational literacy app. We then explore how engagement is related to performance on these games, children's initial literacy skill, and their general progress in reading comprehension across the year. We also examine whether any specific demographic variables are associated with different groups of children.

To identify distinct reader profiles, we draw on analytic techniques used in the analysis of computer games – cluster analysis (Vardal et al., 2022). Cluster analysis, as an unsupervised technique, is well-suited for effective identification of underlying patterns in vast and unstructured datasets. Clustering methods have been used previously to explore academic development and learning by different types of learners (Howard et al., 2018; Liao & Wu, 2023; Lim et al., 2022; Saqr & López-Pernas, 2021; Wang et al., 2023). For example, Sparks et al. (2012) found that the k-means clustering identified three distinct cognitive and achievement profiles of second language learners based on 11 cognitive variables including participants' scores. To date, studies using clustering have typically focused on the application of a single technique, without surveying and comparing various methods available to cluster the data. To address this, we evaluated the application of Gaussian mixture models, partitioning around medoids, clustering large application and hierarchical clustering, alongside the commonly used k-means method.

Our aims are twofold. First, to propose and demonstrate the utility of a novel analytical framework that utilizes the potential offered through rich log files to understand users' behaviour and application in the context of our dataset. Second, to provide guidance about implementation of clustering techniques that can be used to address similar research questions with datasets of similar structure. This work can support additional research objectives, such as understanding how different individuals and cohorts may vary in measured behaviours, for example engagement and performance, over time.

Our research is not driven by a framework of hypothesis testing but is exploratory in nature, driven by the aim to track language and literacy development and identify distinct group of learners. The results have the potential to improve our understanding about what user engagement and performance can tell us about learners and their behaviour. Such investigations also have the potential to inform targeted support for children to foster their development. Our research questions are as follows:

- 1) Do different cluster techniques effectively discern heterogeneous groups, such as distinctive reader cohorts, within massive and unstructured datasets?
- 2) Do the patterns of engagement make sense in relation to theory and previous research?
- 3) Is engagement related to performance?
- 4) Is there a relation between engagement, performance, and student demographics?
- 5) Does the behaviour within and between clusters relate to the programme architecture?

2 Methods

2.1 Research context and participants

The dataset was supplied from a supplemental digital literacy education program - Boost Reading, designed to support the development of core reading skills and concepts to students in kindergarten to Grade 5 through a variety of games. Boost Reading is designed to help children to learn to read (e.g., see https://amplify.com/research-and-case-studies/boost-reading-research/; Newton et al., 2019). It comprises a variety of games, each focusing on different skills that previous research has shown to be critical to the reading development and success (e.g., LARRC & Logan, 2017; Oakhill & Cain, 2012). It uses adaptive technology allowing students to practice the skills they need at their own level and has recommended, but not prescribed, weekly usage. Thus, it allows us to study the development of both word reading and reading comprehension in beginner readers, and the relation between engagement and performance. A different part of the dataset for an independent project has been previously analysed and published by Diprossimo et al. (2023).

The original dataset comprised 43,900 students' behaviour across K to Grade 5. For this paper, we used a subsample of 19,830 students in Grade 2 Elementary schools who played the app from 2020 to 2021. We focused on Grade 2 students to capture the initial stages of reading acquisition, which is a strong predictor of later reading comprehension (Oakhill & Cain, 2012; Verhoeven et al., 2011). Grade 2 students were exposed to a larger variety of games in terms of skill development in the app relative to other grades, allowing for a comprehensive investigation of patterns of user engagement and performance across core skills. The students were located in 347 different schools across 30 different districts, largely in REDACTED FOR REVIEW. A range of indicators available from student log files were used alongside social demographic data to provide insights into engagement and learning across different reading-related skills. A rigorous data cleaning process ensured sufficient data points for the main analyses. We ensured that the final dataset included data only from students who played for a sufficient amount of time to produce meaningful profiles for analyses. We also ensured that the included students had observations spread across the year and across a variety of games to enable meaningful investigation of progress over time whilst ensuring sufficient variation in engagement and performance measures. Our analysis plan was preregistered (https://osf.io/3fhye).

2.2 Data sources

From a theory-driven perspective, our research focuses on four key skills that are essential for reading comprehension: two code-related skills that support word reading, phonological awareness and early decoding, and two meaning-related skills that support reading comprehension, vocabulary and comprehension processes. The app has a hierarchical structure that results in a hierarchical structure of the dataset: skills at the top, followed by games within those skills, levels within the games, attempts, and items within the levels. A total of 30 different games that were played by our sample of Grade 2 users were designed to practice our four target skills, accounting for 54.5% of the total number of games played. We considered data from all of these games: 5 for phonological awareness, 8 for early decoding, 6 for vocabulary, and 11 for comprehension processes. Each game has multiple levels, and a student must complete multiple items within one level to pass to the next. For each level, multiple attempts are allowed. For each attempt, the automatically generated log file provides information on user behaviours and includes: the number of total items attempted, the number of correctly answered items, information on whether the student mastered the level, and scores that are represented by an overall ratio of correct to total items. There are also time related variables such as: the elapsed time and timestamps indicating start and end time for each attempt. These variables are described in the supplementary materials (Table S.1).

It is important to note the mechanism underpinning students' progress within the app. Typically, after mastering a level, pupils are presented with the next level. However, some pupils do not master the level after three attempts. In such instances, that level (and, therefore, the game) is removed from their content for some time. Later, the app puts the student back to the same level of that specific game. Such features not only make each user's journey unique, but also provide personalised training to strengthen each user's skills. Thus, the results of each level and assessment reflect a more realistic personal performance for each user.

An out-of-game assessment, Dynamic Indicators of Basic Early Literacy Skills (DIBELS, 8th edition, University of Oregon, Center on Teaching and Learning, 2018) was administered to capture each child's initial ability at the start of Grade 2. These scores were used to classify students into four performance levels: 'well below benchmark', 'below benchmark', 'at benchmark', and 'above benchmark'. These categories were used to place students at an appropriate starting level in the app.

To provide a measure that captures progress and development of reading comprehension, the app includes a game designed specifically to measure reading comprehension on a monthly basis, a cloze task called Mind the Gap (see supplementary materials Table S.1). The difficulty level of content for this game was designed to be comparable across the year, such that increased scores indicate improved performance. Unlike other games in the app, users have only one attempt (each month), which is time-limited (to three minutes). Scores on this game can be used for overall progress monitoring. Scores on Mind the Gap that either stay the same or improve indicate growth. Performance may also indicate individual disparities in reading ability at the beginning of a student's interaction with the app, allowing us to capture individual differences in trajectories of learning. We used scores on this game as a reference point of users' performance and a validation measure to understand overall learning progress in the app.

2.3 Engagement Indicators

Our choice of engagement indicators was informed by the literature (e.g., Du et al., 2021; Henrie et al, 2015; Sinatra et al., 2015; Yang et al., 2020). We created six game-based indicators to summarise information across usage, considering both the nature of the game and the data, while avoiding redundancy: *all hours spent, variety of games played, all attempts, the number of levels played, days of playing,* and *proportion of early exit rounds.* These are listed in the appendices (Table A.1). These variables were fed into our exploratory analyses to explore variation across various skills, games, and users themselves.

3 Data analysis

We summarise the stages involved in the data analysis in Figure 1. Each step is detailed below, and will be referred to in the data analysis and results sections.

Data pre-processing	Step 1 Data pre-processing	Amplify Reading log file data 19,830 users in Grade 2 2020-2021	
		Integrate with DIBELS and social demographics of each user	
		Construct 6 engagement variables for each user across four skills PA: phonological awareness ED: early decoding V: vocabulary CP: comprehension processes	Engagement Variables • _all hours spent_ on PA/ED/V/CP • _variety of games played_ within PA/ED/V/CP • _all attempts_ within PA/ED/V/CP • _the number of levels played_ within PA/ED/V/CP • _days of playing_ within PA/ED/V/CP • _proportion of early exit rounds_ within PA/ED/V/CP
		Use engagement information to filter out students with no engagement in four skills, remaining with 19,305 users	
		Standardise engagement variables	
Application of clustering methods	Step 2 Apply algorithms	Apply 5 clustering algorithms to 24 engagement variables	Clustering Algorithms • Gaussian mixture models • k-means • k-medoids (partition around medoids) • clustering large applications • hierarchical clustering
	Step 3 Evaluate algorithms	Evaluate the clustering results using cluster validation metrics	Metrics • silhouette coefficient • Dunn index • within clusters sum of squares
	Step 4 Decide on the optimal algorithm for the data	Use visualization of the clustering algorithms results and validation metrics to select the most appropriate method for the data	
Data processing after application of the method	Step 5 Link the results to social demographics and DIBELS	Order clustering results by initial ability (DIBELS)	Social Demographics • gender • English language learner • special education needs • race/ethnicity
	Step 6 Add in-game performance	Use performance data to construct 4 performance variables for each skill and integrate that with clustering results *Mastery defined as the equivalent to "passing" the game level by achieving the score equal or above the threshold of master.	Performance Variables • _attempts to mastery_* for a level within PA/ED/V/CP • _number of levels mastered_ within PA/ED/V/CP • _speed to mastery_ for a level within PA/ED/V/CP • _time to mastery_ for a level within PA/ED/V/CP
Presentation of clustering results	Step 7 Present engagement and performance of clusters	Visualise engagement and performance of each cluster using radar charts	Generate Variable • ranking for each variable of engagement and performance
	Step 8 Validate clustering results using out-of-game performance	Validate clustering results by relating to DIBELS and social demographics (both visually and using descriptive tables)	
		Validate clustering results by relating to Mind the Gap progress	Mind the Gap progress = Mind the Gap score (average) at the end – at the start

Figure 1. 8 Steps of data analysis for the two aims of our study.

The clustering analyses aggregated performance across the entire year of activity. All statistical analyses were conducted using openly available software R (R Core Team, 2021). We utilized the factoextra package to perform partitioning clustering, hierarchical clustering, as well as visualization and silhouette coefficients of the clustering results (Kassambara & Mundt, 2016). To determine the optimal number of clusters, we employed the mclust package and examined the Bayesian Information Criterion (BIC) results generated by model-based clustering with various cluster numbers (Fraley & Raftery, 1998, 2006). R package fpc was used to compute two metrics for cluster validation – Dunn index and within clusters sum of squares (Hennig, 2023). Data pre-processing was conducted using tidyverse and dplyr packages (Wickham et al., 2019a, 2019b).

3.1 Step 1: Data pre-processing

Our steps to ensure data quality are outlined in Figure 1, Step 1. We constructed indicators of engagement in the app using the available data. To homogenise the size and variability of these input variables, the built-in-R function "scale" was used for standardising across the six engagement variables, involving the conversion each original value into a z-score (Milligan & Cooper, 1988).

3.2 Step 2: Cluster algorithms

To select the best model for our data and the optimal number of clusters *K*, we used Gaussian mixture models and the Bayesian Information Criterion (BIC; Fraley & Raftery, 1998), where higher values of BIC indicate better fit (Bozdogan, 1987; Kvapil et al., 2022; Schwarz, 1978). Once the optimal number of clusters was determined, we compared the following widely-used clustering methods: (1) Model-based clustering analysis (Gaussian mixture models); (2) Partition clustering (k-means, k-medoids, clustering large applications);

and (3) hierarchical clustering (see Banfield & Raftery, 1993; Hastie et al., 2009; Kaufman & Rousseeuw, 1990; MacQueen,1967).

3.2.1 Model-based clustering analysis (Gaussian mixture models)

In this approach, data are viewed as coming from a mixture of probability distributions, each representing a different cluster. The same object could be assigned to more than one cluster by Gaussian mixture models. Hence it is a soft clustering algorithm, distinct from hard clustering algorithms such as partition clustering (e.g., k-means, k-medoids and CLARA) and hierarchical clustering. Each cluster is modelled by the normal or Gaussian distribution which is characterised by the parameters (e.g., mean and variance).

3.2.2 Partition clustering

Partition clustering classifies objects into multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible and objects from different clusters are as dissimilar as possible. A partitioning algorithm divides a dataset into a set of disjoint partitions, with each partition representing each cluster. We used k-means, k-medoids and clustering large applications based on this approach.

3.2.2.1 K-means

K-means works by assigning the mean of points to the centre of each cluster. The results may be sensitive to the initial random selection of cluster centres, because the choice of the distinct set of initial centres will lead to different clustering results on different runs of the algorithm. K-means clustering is the most widely-used unsupervised machine learning algorithm due to its simple and efficient features and has been used previously as a clustering method in digital learning environments (Liao & Wu, 2023; Moubayed et al., 2020).

3.2.2.2 K-medoids (partition around medoids)

K-medoids clustering is considered more robust than k-means (Park & Jun, 2009). Each cluster is represented by a selected object within the cluster, which is the most centrally located point in each cluster: a medoid. Clusters are defined so that the medoid (an object) within a cluster is one for which the average dissimilarity between itself and all the other points is minimal. The partition around medoids (PAM) algorithm is the most common k-medoids clustering method. *K* representative medoids among the observations are identified, clusters are constructed by allocating the remaining observations to the nearest medoids. Next, the selected object (medoids) and non-selected objects are swapped to improve the quality of clustering by minimizing the sum of the dissimilarities of all objects to their closest representative object.

3.2.2.3 Clustering large applications (CLARA)

To address the computational storage challenges posed by the PAM algorithm for large datasets, clustering large applications (CLARA) can be considered. The difference between the algorithms is that PAM searches for medoids within the entire dataset, whilst CLARA considers a small, fixed size sample of the dataset and applies the PAM algorithm to generate the optimal medoids for the sample. To reduce sampling bias, CLARA repeats the clustering and sampling procedures a pre-specified number of times. The set of medoids with the minimal objective function is represented by the final clustering results. The objective function corresponds to the sum of the dissimilarities of the observations to their nearest medoids, which is also a measure of the goodness of the clustering.

3.2.3 Hierarchical clustering

Hierarchical clustering divides a dataset into a sequence of nested partitions. Agglomerative clustering is the most common type of hierarchical clustering and works in a "bottom-up" manner. The algorithm starts by treating each object as a single-element cluster (leaf), then the most similar pairs of clusters are successively combined into a bigger cluster using the linkage function until all clusters have been merged into one big cluster (root) containing all objects.

3.3 Step 3: Cluster validation

Although the outlined clustering methods may appear quite different in terms of their underlying mechanisms, a variety of validation metrics can be used to compare across all to assess the goodness of the clustering algorithm results. We used three quantitative evaluations to assess cluster validity and provide indications of how well the clustering fits the data: silhouette width, Dunn index, and within cluster sum of squares (Halkidi et.al, 2015; Theodoridis & Koutroumbas, 2006). Selection of the best clustering solution optimised simultaneously maximal within-group homogeneity while maintaining meaningful between group heterogeneity or, in simple words, ensuring that resulting clusters were as distinct as possible from each other. It is not uncommon for clustering results to be sensitive to data structure and initiation data points. Thus, to complement numerical metrics we also used insights from the field to guide the selection of the best clustering solution considering what we know about expected patterns of associations between engagement and performance.

4 Results

4.1 Sample characteristics & engagement descriptive statistics

The initial DIBELS results indicated that most users in the sample were in ‘well below benchmark’ (40%) and ‘at benchmark’ initial reading ability groups (27%). The proportion of males and females was balanced. Most of the children were non-English language learners (non-ELL) (62%) and non-special education needs (non-SEN) students (81%). The highest proportion of races were Hispanic or Latino (64%), followed by Black or

African American (9%), with Alaskan Native being the lowest (0.03%). The full demographic descriptive statistics are reported in the appendices (Table A.2).

The engagement data revealed a notable emphasis on games that targeted two skills: early decoding and comprehension processes, as indicated by higher mean values for *all hours spent*, *variety of games played*, and *all attempts*. However, games within these two skills showed a relatively high *proportion of early exit rounds* (22% and 23% on average, respectively) compared to 16% for games within phonological awareness, indicating a tendency for students to end their attempts prematurely. The games within early decoding and comprehension processes were frequently played by students throughout the year, evidenced by high values for *days of playing* (9 and 7 days on average, respectively) and *the number of levels played* (10 and 8 levels on average, respectively). Of note, vocabulary games had lower time requirements, possibly indicating a more relatively easygoing or shorter gameplay experience. Students had least exposure to games in the phonological awareness skill. The full descriptive statistics of engagement data are provided in the supplementary materials (Table S.2).

4.2 Step 4: Results of clustering algorithms

4.2.1 Number of clusters

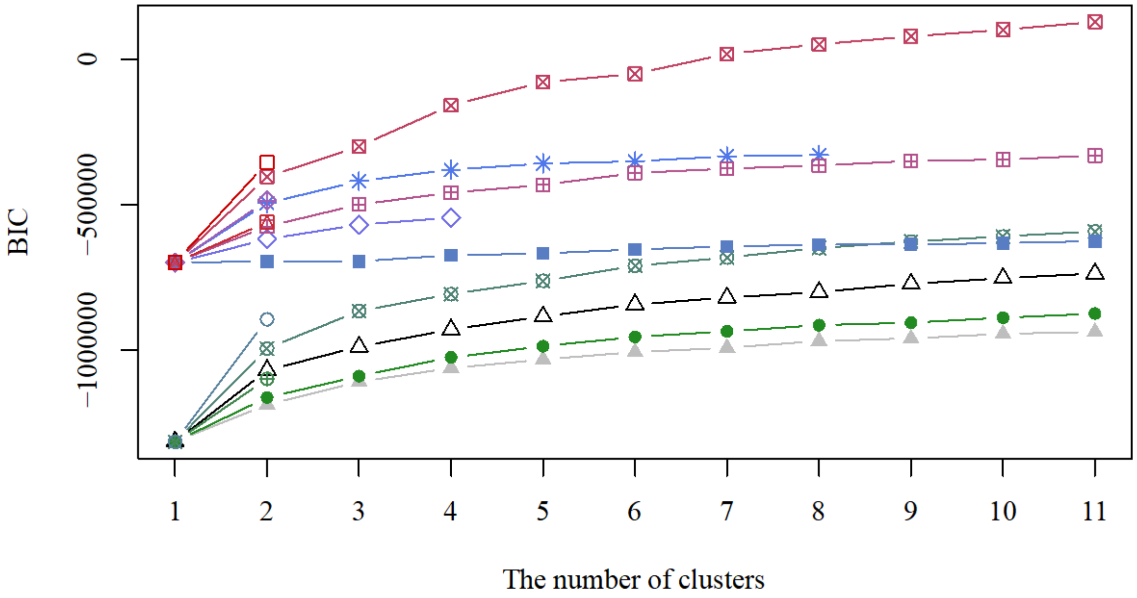
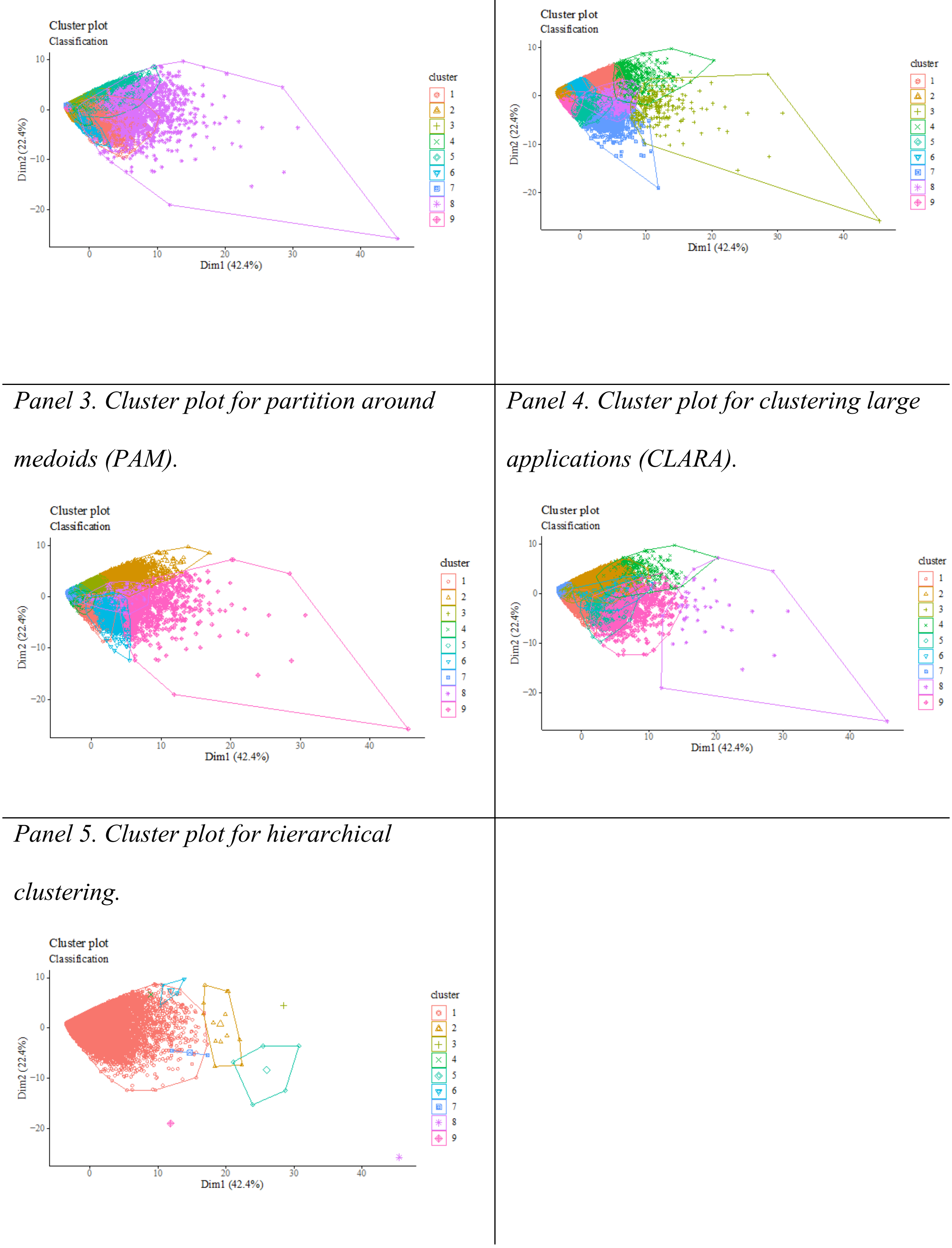


Figure 2. The Bayesian Information Criterion (BIC) for model-based methods (Gaussian mixture models) applied to the engagement data with optimal clusters.

The BIC curve remained relatively flat beyond nine clusters indicating nine clusters with different levels of engagement as the best solution (Figure 2). We used this number of clusters across the five algorithms we compared. The principal component analysis approach was used to visualize the clustering results in two dimensions. Visualization of the clustering result was plotted with each point representing one student, grouped into different clusters (Figure 3).



Panel 1. Cluster plot for Gaussian mixture model (GMM).

Panel 2. Cluster plot for k-means.

Panel 3. Cluster plot for partition around medoids (PAM).

Panel 4. Cluster plot for clustering large applications (CLARA).

Panel 5. Cluster plot for hierarchical clustering.

Figure 3. Visualization of five clustering results on a scatter chart (two dimensions).

From visual exploration (Figure 3), we can see that the Gaussian mixture model performed less well than other methods, since the data points were not well separated (Panel 1). As shown in Panels 2, 3, and 4, CLARA tended to group some values at the extremes (purple points: Cluster 9) into one group, while the group containing these extreme values was extended to include more centre points by k-means (green points: Cluster 3) and PAM (pink points: Cluster 9). The results of the hierarchical clustering algorithm were the weakest; this method could not separate the data into distinct groups, evident from large overlaps across most of the data points in the plot (Panel 5).

4.2.2 Validation indices for cluster algorithms

The results of clustering evaluation based on internal validation for our five clustering methods are provided in Table 1. Indicators used to compare methods were silhouette coefficient and Dunn index, where higher values indicate a better fit, and within clusters sum of squares (WCSS), where lower values indicate a better solution. Cohort size and the average silhouette width for each cohort for each method are reported in supplementary materials (Table S.3).

Clustering methods	*Internal validation metrics*		
	Silhouette coefficient	Dunn index	Within clusters sum of squares (WCSS)
Gaussian mixture models (GMM)	-0.03	0.001	262873.5
K-means	0.17	0.001	181402.3

K-medoids (partition around medoids (PAM))	0.13	0.002	194398.9
Clustering large applications (CLARA)	0.09	0.001	202077.8
Hierarchical clustering	0.67	0.213	443703.3

Table 1. The validation results using three internal validation metrics for five clustering methods.

Although hierarchical clustering had the highest silhouette coefficient and Dunn Index value, it was discarded due to identifying only one sample of adequate size (19277 = 99.85% of sample). The lack of separation between clusters was confirmed by the high value of WCSS. The next two best methods were k-means and PAM, with k-means outperforming PAM on two metrics: silhouette coefficient and WCSS (see Table 1). The Dunn index metric did not reliably distinguish between these two methods (all values between 0.001 - 0.002 apart from hierarchical clustering). Given this comparison, k-means was found to be optimal because there was maximal within-group homogeneity with meaningful between group heterogeneity, supported by two of three internal validation metrics. Therefore, this solution was used for the remainder of our analysis.

4.3 The clustering results obtained through k-means

4.3.1 Clustering results of engagement levels

Clustering of the six engagement variables identified nine distinct groups representing different levels of participation. Each engagement indicator was formed using the mean for each cluster which was then used as an input in ranking from best to worst to characterise these nine clusters (see appendices, Table A.3). Using this approach, consistently similar rankings were observed among five of the six indicators: *all hours spent, variety of games played, all attempts, the number of levels played* and *days of playing* for each group. The exception was *proportion of early exit rounds* which showed the opposite ranking. Students who invested more time in a skill tended to play more frequently, explored a more diverse set of games, encountered more levels, and made more attempts; as a result, these students had a low proportion of early exists from their attempts.

Furthermore, diverse student profiles exhibited varying levels of engagement across skills, indicating unique experiences and skill 'preferences'[1] across users. For example, some groups showed low engagement across four skills (e.g., Cluster 4), some showed high engagement (e.g., Cluster 5), with others displayed clear skill 'preferences'. In line with the overall design of the skills, individuals who engaged more in code-related phonological awareness games also tended to play another set of code-related games - early decoding - more frequently; in contrast, those who participated more in meaning-related vocabulary games also showed higher involvement in games targeting comprehension processes. The clusters identified students with similar game experiences, as well as identifying nine distinct groups with potentially diverse experiences, enabling exploration of the impact of effort invested in one skill on other skills.

4.3.2 Step 5: Initial literacy performance of the clusters

To aid interpretation, the clusters (see Figure 4) were ordered by the initial literacy assessment performance (DIBELS): from the cluster with the lowest proportion of students who were well below benchmark (Cluster 1) to the cluster with the highest proportion of

[1] Students can't select skills, games, or levels. Skill 'preferences' reflect their most engaged skill within the app.

students who performed well below benchmark (Cluster 9). The dashed lines represent the 50% and 75% thresholds for each cluster.

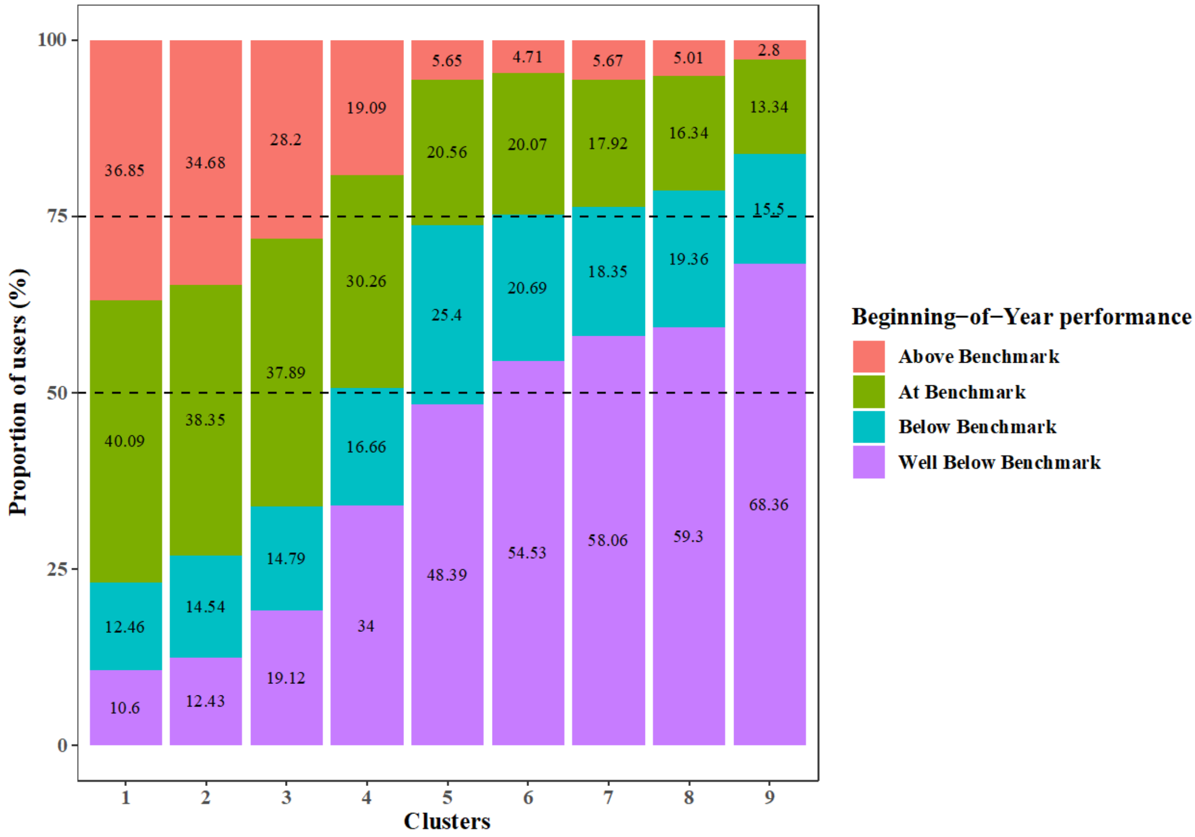


Figure 4. Proportion of children in each cluster performing at different levels of DIBELS at beginning of year.

The majority of students in the Clusters 1-3 performed at or above DIBELS benchmark. In Cluster 4, approximately half of the students met or exceeded the benchmark. From Clusters 5 to 9, approximately 75% or more of students performed below the benchmark level. Table 3 shows the number of participants in each cluster alongside key demographics (for more detailed social demographics information see supplementary materials, Table S.4, and section 4.3.4.1).

4.3.3 Step 6 & 7: Visualizing engagement and performance levels of each cluster through radar charts

After implementing the cluster algorithms, we calculated the total number of attempts and the time required for a student to master a specific level, excluding incomplete attempts, which indicated that a student had made an early exit from the game.

	Q1	*Median*	*Mean*	*Q3*	*SD*	*68% lies between*	*95% lies between*
Phonological Awareness	2.13	3.86	6.86	8.46	8.50	(0, 15.36)	(0, 23.86)
Early Decoding	2.37	3.47	5.76	6.06	7. 85	(0, 13.61)	(0, 21.45)
Vocabulary	1.39	2.18	3.74	3.96	5.66	(0, 9.40)	(0, 15.05)
Comprehension Processes	2.16	3.41	5.47	6.14	6.63	(0, 12.09)	(0, 18.72)

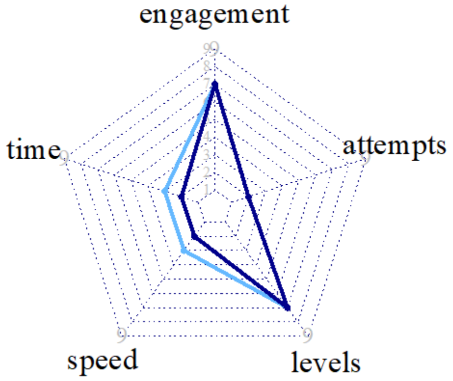
Table 2. A comparison of the indicators of the time taken to succeed for games that target the four different skills.

Table 2 provides the descriptive statistics for one of the four performance indicators: time to mastery, which represents the duration from the first attempt to the first mastery. This indicator is a key measure across various levels within a specific game and also links directly the elapsed time within attempts and the number of attempts. The vocabulary games were the fastest to master, with least variation, whilst the phonological awareness games required the longest time and show the highest variability. The average times to master a level in the early decoding and comprehension processes games were comparable.

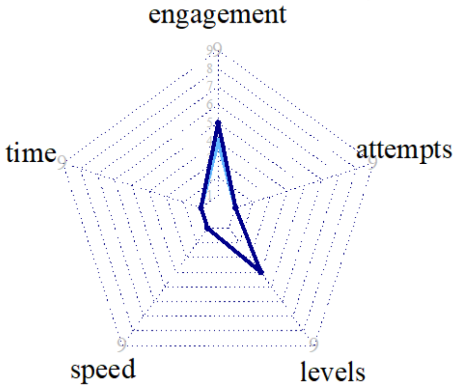
Visualizations to illustrate the diverse outcomes of student engagement across the four skills were created from the average ranking of the five consistent engagement variables obtained in Step 4 (as shown in appendices, Table A.3), alongside the four indicators evaluating performance on each skill: *attempts to mastery, number of levels mastered, speed to mastery,* and *time to mastery.* Both engagement and performance variables were ranked in descending order from indicators of best to worst engagement/performance, with the highest (or best) rank displayed as closest to the centre in the charts (for more details, see supplementary materials, Table S.5). We first discuss the analyses of the code-related skills (Figure 5), followed by the meaning-related skills (Figure 6).

4.3.3.1 Code-related skills: Phonological awareness and early decoding

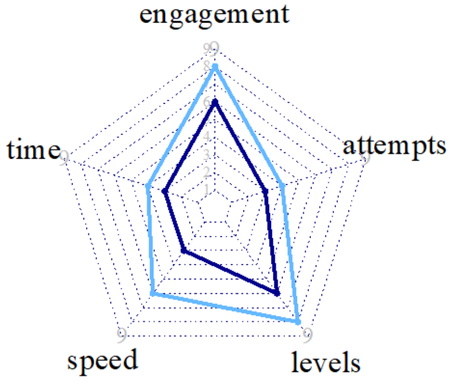
Cluster 1 (n = 1726, 8.94%)



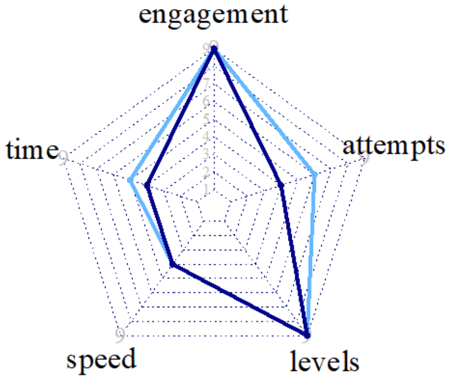
Cluster 2 (n = 571, 2.96%)



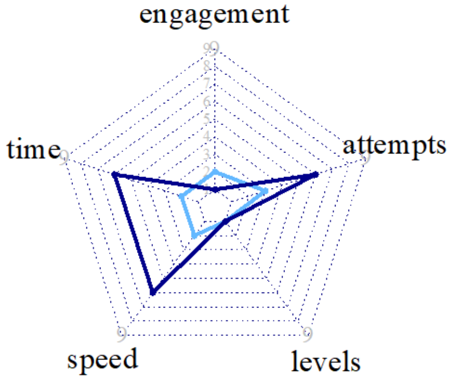
Cluster 3 (n = 3468, 17.96%)



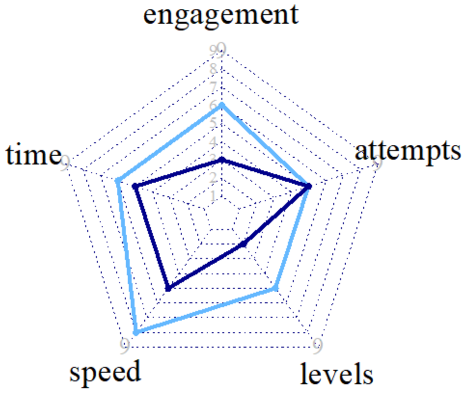
Cluster 4 (n = 4862, 25.19%)



Cluster 5 (n = 248, 1.28%)



Cluster 6 (n = 1126, 5.83%)



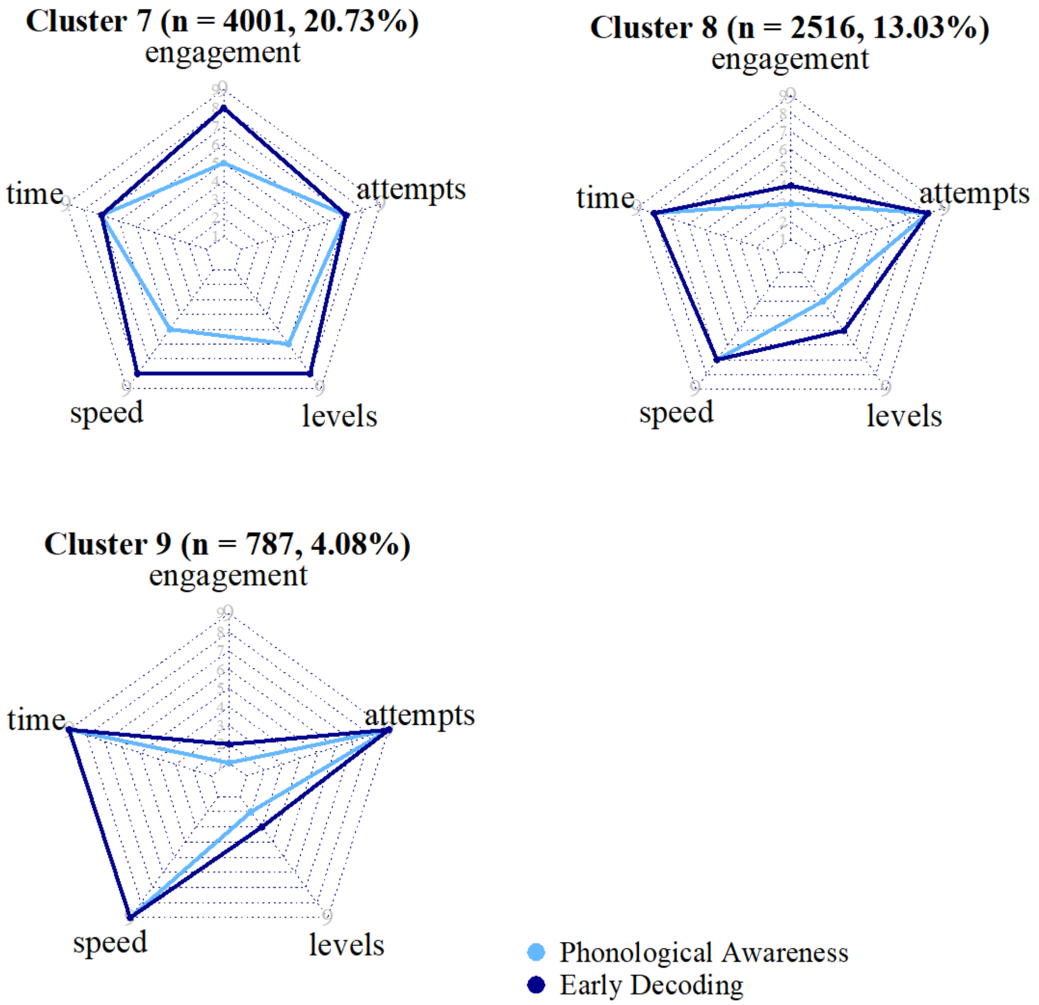


Figure 5. Radar Chart: Engagement and performance rankings in phonological awareness games (light blue) and early decoding games (dark blue) among nine clusters.

We focus on the clusters with the largest number of participants to highlight meaningful differences: Clusters 3, 4, and 7 (17.96%, 25.19%, and 20.73% of participants). These groups differed in initial DIBELS literacy skills, but each showed low to medium engagement in both phonological awareness and early decoding. This finding in line with the low rankings for *number of levels mastered* in phonological awareness and early decoding for these three groups. The discrepancy in the time and effort invested by these groups in both phonological awareness and early decoding is minimal (see Table S.6, supplementary materials); they mastered more early decoding games than phonological awareness games.

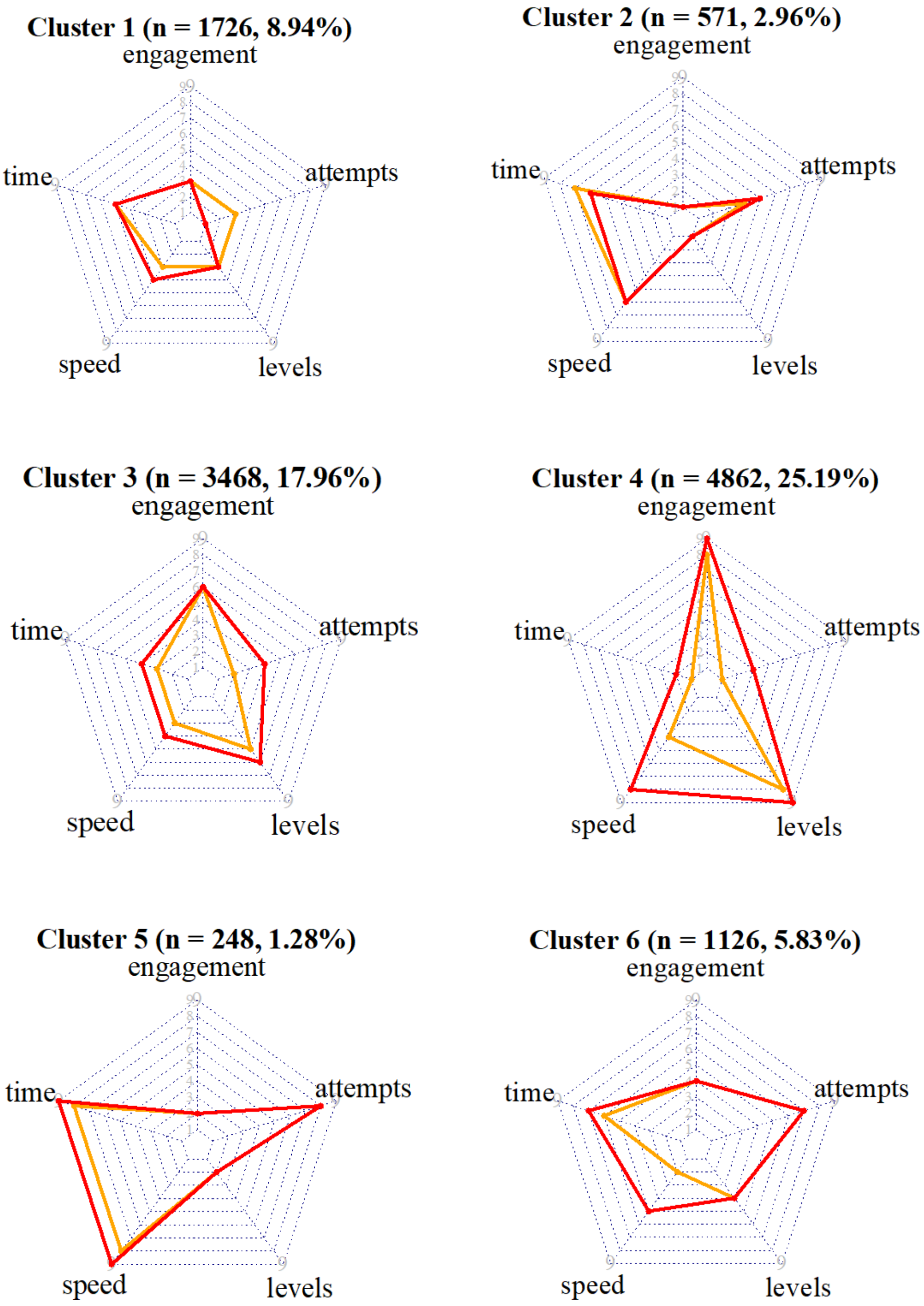
This could be attributed to the longer total duration spent on early decoding compared to phonological awareness.

Critical differences among these three groups are evident when we look at the indicators of engagement together with performance on phonological awareness and early decoding games. Cluster 7, which had the lowest literacy levels of these three groups, was more highly ranked on engagement in phonological awareness compared with the other two groups, yet their performance was ranked similarly to the other two groups (Figure 5 & Table S.7) on two metrics: (1) the average number of levels mastered per hour (Cluster 7 = 6.7, Cluster 4 = 6.9, Cluster 3 = 5.4 levels); and (2) the average time required to succeed (Cluster 7 = 7.1, Cluster 4 = 6.2, Cluster 3 = 5.6 mins). Although the difference in time to mastery difference exceeds one minute, children in Cluster 7 engaged with a larger number of levels suggesting that they might perceive the level to be challenging.

Similarly, for early decoding, these three groups had comparable levels of engagement but varied in performance. Clusters 3 and 4 had similar rankings (3rd and 4th respectively) for time and attempts required to master a level, whereas Cluster 7 was comparatively less proficient and ranked 7th (Figure 5). Exploratory data analysis revealed that Cluster 7 predominantly engaged with the initial foundational levels of the game (see supplementary materials, Figure S.8). From the radar chart, the ranking of Cluster 7 towards the outer edges in terms of *times to mastery, attempts to mastery* and *speed to mastery* indicates that the difficulty they faced in the early decoding games, resulted in needing more time than others to master even the initial levels.

Clusters 1-4 had lower engagement with phonological awareness and early decoding games, but showed better performance in terms of *time to mastery* and *attempts to mastery* (Figure 5). Conversely, Clusters 6-9 exhibited higher engagement in these two skills but demonstrated weaker *time to mastery* and *attempts to mastery*. Regardless of group, the ranking pattern for *time to mastery* and *attempts to mastery* remained consistent in phonological awareness and early decoding games.

4.3.3.2 Meaning-related skills: Vocabulary and comprehension processes


Cluster 1 (n = 1726, 8.94%)

Cluster 2 (n = 571, 2.96%)

Cluster 3 (n = 3468, 17.96%)

Cluster 4 (n = 4862, 25.19%)

Cluster 5 (n = 248, 1.28%)

Cluster 6 (n = 1126, 5.83%)

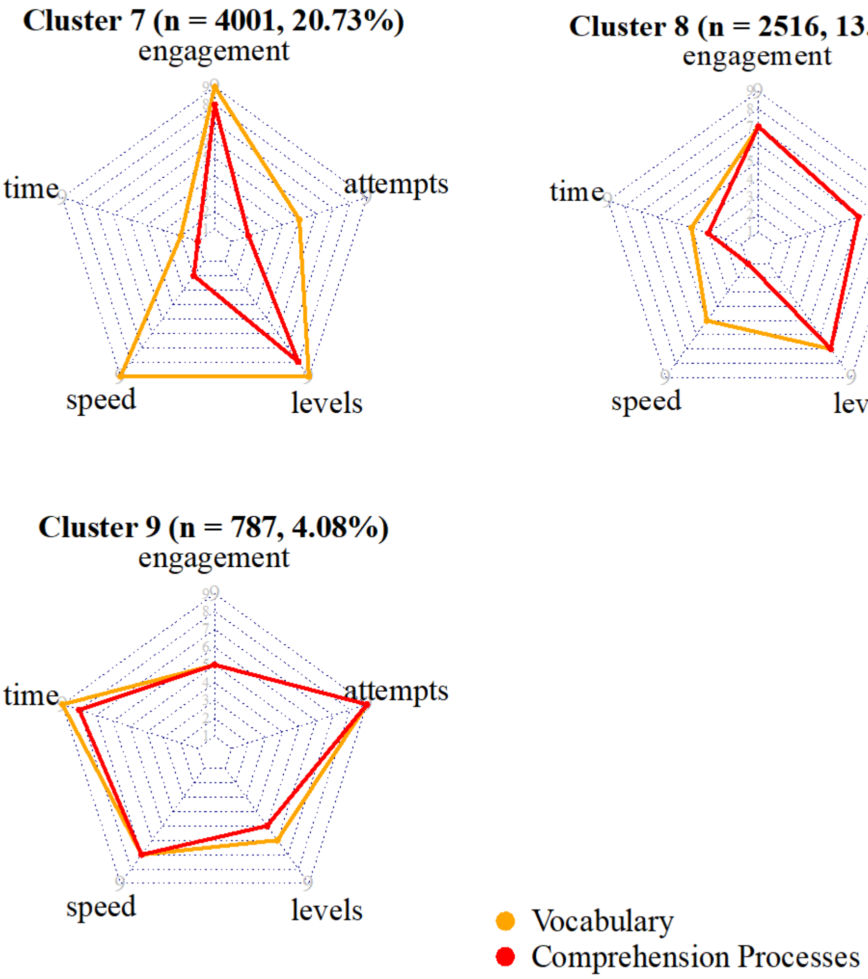


Figure 6. Radar Chart: Engagement and performance in vocabulary games (orange) and comprehension processes games (red) among nine clusters.

For vocabulary games, Cluster 3 exhibited the highest participation level compared with Clusters 4 and 7 and, on average, mastered 5-6 more levels (supplementary materials, Table S.7). The exploratory data analysis revealed that a notably higher number of students in Cluster 3 mastered over 20 levels compared to the other groups (supplementary materials, Figure S.9). Cluster 3 performed well in the vocabulary game, achieved the highest number of levels mastered per hour, but their *time to mastery* and *attempts to mastery* for each level were similar to the other groups. These findings suggest that the proficiency level of vocabulary games had minimal impact on the time requirements for mastery. For comprehension processes games, Cluster 3 also demonstrated the highest level of participation. Based on the slightly higher number of levels played, this group required slightly more time and attempts to master each level compared to the other two groups.

The overall engagement levels depicted in the radar charts (Figure 6) show that Clusters 1-3 had the highest participation in vocabulary and comprehension processes games, with a close ranking association between their *time to mastery* and *attempts to mastery*. In contrast, Clusters 7-9 had lower engagement and higher *attempts to mastery*, indicating these skills posed a challenge for the users in these groups.

4.3.4 Step 8: Validation of clustering results

4.3.4.1 Relation to demographics and initial literacy ability

Different patterns of engagement were evident, even for clusters with similar literacy skills. Here we discuss the clustering results in relation to demographics and DIBELS. To aid interpretation, we consider pairs, triplets, or quadruplets of groups with similar initial literacy skills. Of note, the sizes of Clusters 2, 5, and 9 were relatively small in relation to the other clusters, Cluster 2 had a greater proportion of male students, Cluster 8 a greater proportion of female students, and Clusters 5-9 had greater proportions of ELL and SEN students (Table 3).

Cluster (n = cohort size, %)	Gender		English Language Learner - ELL (%)	Special Education Needs - SEN (%)	DIBELS			
	Male (%)	Female (%)			Well Below Benchmark (%)	Below Benchmark (%)	At Benchmark (%)	Above Benchmark (%)
Cluster 1 *n = 1726, 8.94%*	48.96	46.06	10.72	4.69	10.6	12.46	40.09	36.85
Cluster 2 *n = 571, 2.96%*	54.47	41.86	18.04	6.48	12.43	14.54	38.35	34.68
Cluster 3 *n = 3468, 17.96%*	49.28	46.14	15.46	4.64	19.12	14.79	37.89	28.2
Cluster 4 *N = 4862, 5.19%*	46.24	45.99	23.51	6.09	34	16.66	30.26	19.09

Cluster 5 *n = 248, 1.28%*	47.58	44.35	39.92	10.89	48.39	25.4	20.56	5.65
Cluster 6 *N = 1126, 5.83%*	47.42	44.14	25.75	8.53	54.53	20.69	20.07	4.71
Cluster 7 *n = 4001, 20.73%*	43.74	48.16	35.99	11.00	58.06	18.35	17.92	5.67
Cluster 8 *n = 2516, 13.03%*	42.45	50.48	32.39	10.37	59.3	19.36	16.34	5.01
Cluster 9 *n = 787, 4.08%*	49.05	45.11	38.37	14.23	68.36	15.5	13.34	2.8

Table 3. Diverse social demographics of students in nine clusters. Social demographics were not provided for some individuals, resulting in missing data.

Initial literacy performance of Clusters 1-3 was high with 66% or more at or above DIBELS benchmark; Cluster 3 was the weakest. Cluster 2 was a small cluster, with a greater proportion of students reported as ELL and those with SEN than Clusters 1 and 2. In relation to engagement and performance, Cluster 2 showed higher engagement with the meaning-related games (vocabulary and comprehension processes), but a greater number of attempts to master a level, Cluster 1 engaged more with these games and performed less well than Cluster 3.

Cluster 4 resembled Clusters 1-3 with approximately half the group at DIBELS benchmark. Cluster 4 was the largest group, had balanced gender ratio, and a greater proportion of ELL students compared with Clusters 1-3. The proportion of students with reported SEN in Cluster 4 was similar to that of Cluster 2. Children in Cluster 4 exhibited minimal participation in any of the games studied. Compared with Clusters 1-3, they took longer time and needed more attempts to master a level in code-related games, but they outperformed Clusters 1-3 in meaning-related games.

Clusters 5-7 exhibited similar DIBELS performance to each other with fewer than 25% of individuals achieving benchmark. In comparison to Clusters 1-4, Clusters 5-7 had a higher proportion of ELL students. Within this set of groups, Cluster 6 had the lowest proportion of students as ELL or with SEN, Cluster 5 had the highest proportion of ELL students, and Cluster 7 the highest proportion of students with SEN. Cluster 5 demonstrated the highest level of engagement with both code-related and meaning-related games, and excelled particularly in code-related games. Cluster 6 demonstrated better performance in meaning-related games compared with Cluster 5.

Turning to those with the weakest initial literacy performance, Cluster 9 performed more poorly than Cluster 8. Cluster 9 had a slightly higher proportion of students who were at-risk of poor literacy (ELL and those with SEN) relative to Cluster 8. Cluster 8 outperformed Cluster 9 in terms of engagement and performance, particularly in meaning-related games, but Cluster 9 had higher levels of participation compared with Cluster 8.

4.3.4.2 Relation to progress on Mind the Gap

Finally, we relate the findings on engagement to general performance and learning that is captured by the monthly progress monitoring game - Mind the Gap (see Table 4 for gain in scores across the year, and the number of times this game was completed). Starting with Clusters 1-3, who had the highest proportion of students with good initial literacy scores, progress of students in Cluster 2 is of note: they engaged most with the app, made the most progress, but also had a higher proportion of students known to be at-risk of reading difficulties (ELL; students with SEN). Clusters 4 and 7 made little progress and showed the least engagement across the school year. Thus, engagement with Mind the Gap validates the clustering results produced by k-means using six indicators of engagement of skills. An exception to this pattern is evident for Clusters 8 and 9: neither made substantial positive progress, despite regular engagement.

Cluster *(n = cohort size, %)*	*Average of start scores (SD)*	*Average of difference (progress) [end - start] (SD)*	*Number of times took Mind the Gap (SD)*
Cluster 1 *n = 1726, 8.94%*	8.25 (7.04)	1.09 (8.42)	8.05 (2.39)
Cluster 2 *n = 571, 2.96%*	7.28 (7.53)	1.34 (8.44)	9.22 (2.37)
Cluster 3 *n = 3468, 17.96%*	6.79 (7.14)	0.92 (7.69)	5.72 (2.44)
Cluster 4	5.61 (6.88)	0.58 (6.14)	2.83 (1.82)

n = 4862, 25.19%			
Cluster 5 *n = 248, 1.28%*	1.92 (4.34)	0.41 (6.86)	9.33 (2.22)
Cluster 6 *n = 1126, 5.83%*	2.08 (4.94)	0.59 (6.78)	8.01 (2.39)
Cluster 7 *n = 4001, 20.73%*	2.50 (5.58)	0.03 (6.06)	3.78 (2.16)
Cluster 8 *n = 2516, 13.03%*	2.27 (5.33)	0.04 (6.74)	6.49 (2.39)
Cluster 9 *n = 787, 4.08%*	1.42 (4.78)	-0.20 (6.25)	8.12 (2.27)

Table 4. Mind the Gap (out-of-game assessment) results for each cohort from 2020 to 2021.

5 Discussion

This innovative study provides critical information about the use and evaluation of cluster analysis to understand student engagement and usage with an educational literacy app in relation to learning outcomes. It provides a novel approach by integrating data analysis, data cleaning, methods, and validations into a holistic framework. Here we highlight key strengths of the study and focus on contributions made to methodology, what these findings tell us about literacy development, and also limitations and suggestions for future work.

5.1 Do different cluster techniques effectively discern heterogeneous groups, such as distinctive reader cohorts, within massive and unstructured datasets?

All clustering methods performed well except for hierarchical clustering, and model comparison identified that the k-means algorithm produced the best clustering performance. In line with previous research (Theodoridis & Koutroumbas, 2006; Yoo, 2020), k-means demonstrated sufficient dimension reduction and easily clustered high-dimensional numerical data. One crucial consideration in clustering analysis is the dimensionality and nature of the dataset, which will influence the similarity measures for different algorithms (Shirkhorshidi et al., 2015). We found that hierarchical clustering was less effective in classifying numerical data when compared to the partition clustering. However, other work has shown its effectiveness for the clustering of categorical data (Šulc & Řezanková, 2019). When comparing the three methods of partition clustering, partitioning around medoids and clustering large applications demonstrated notably faster execution times compared to k-means (see also Arora et al., 2016; Kaufman & Rousseeuw, 1990). Overall, when considering the research questions and nature of the data, it is essential to select an appropriate clustering method for a given dataset.

The use of radar charts effectively captured a wealth of information to show the association between various indicators of engagement and performance separately for each of the nine clusters. Examination and comparison of the size and shape of the radar chart for each cluster confirmed differences between reader groups that were validated by other metrics, which we discuss below.

5.2 Literacy development

5.2.1 Do the patterns of engagement relate to theory and previous research?

We mapped the engagement (and also performance) data for the four skills in the radar charts, plotted separately into two pairings: code-related skills (phonological awareness and early decoding), and meaning-related skills (vocabulary and comprehension processes). Separate plots were constructed for each cluster. When we examined the patterns of engagement, we found validation that our clustering algorithm identified coherent and meaningful groupings in relation to both theory and empirical research. Note, that these theoretically-informed pairings were not included into the data fed into the clustering algorithms.

First, when we look within clusters, high engagement with phonological awareness games was found alongside high engagement with early decoding games. Similarly, high engagement with vocabulary games was associated with high engagement with games targeting comprehension processes. This pattern is in line with empirical research that has demonstrated that different skills are associated with word reading and text comprehension (Kendeou et al., 2008; Muter et al., 1998; Oakhill & Cain, 2012). Second, comparison across clusters indicated that groups who had high engagement with code-related skill games tended to show low engagement with meaning-related skill games, and vice versa. This is in line with a developmental shift from word reading mastery to listening comprehension in the determination of reading comprehension (Gough & Tunmer, 1986; LARRC, 2015). Third, when we consider these patterns in relation to initial reading ability, we have further validation that the clusters are meaningful. Students with lower initial reading ability tended to show higher engagement in basic code-related skills, whereas students with higher initial reading ability tend to exhibit higher engagement in meaning-related skills.

5.2.2 *Is engagement related to performance?*

Students who demonstrated moderate to high engagement levels in basic code-related skills were more likely to demonstrate lower to moderate performance in these specific games. Our data strongly suggest that students' varying levels of participation can partially account for the variation in their progress and reading outcomes in line with Ericsson's work on the importance of practice (1993). Our approach to quantifying student performance involved calculating the average time taken by each student within a group to master each level across the school year. A strength of this approach is that we utilized time to succeed for each student, allowing us to obtain a comprehensive measure of central tendency. However, a limitation is that this can be influenced by other students in the group so may not fully reflect individual progress during the year. Students who had greater exposure to phonological awareness and early decoding games, may appear slower in terms of the average values of time to mastery, but it is still possible for them to benefit from the related games and make progress through engagement. Students who shown greater engagement in vocabulary and comprehension processes games, tended to exhibit fewer attempts but required more time to master each level. This may reflect the fact that reading for meaning requires greater cognitive effort than word reading, due to the need to draw on both decoding and language comprehension skills.

Most students demonstrated good engagement with the app and positive progress on the out-of-game assessment, Mind the Gap. In line with a growing body of research into digital educational technology and literacy (see Biancarosa & Griffiths, 2012; Diprossimo et al., 2023; Nizam and Law, 2021; Yang et al., 2020), our findings suggest that incorporating digital reading supplements in classroom-based activities offers promising opportunities for early literacy education.

5.2.3 Is there a relation between student behaviour and demographics?

The groups with lower literacy ability and higher engagement in the code-related skills had a higher proportion of students reported as ELL and SEN. This finding aligns with previous work showing that these populations often have literacy difficulties (Fitzgerald, 1995; Melby-Lervåg & Lervåg, 2014). We note that those groups who had a high proportion of at-risk students who engaged regularly with the program benefited more. For example, Clusters 7 and 8 each had a high proportion of students reported as ELL or SEN (and comparable initial low literacy levels), but Cluster 8 had higher engagement and made more progress. This suggests that engagement with the programme can benefit students who are at-risk of poor literacy.

5.2.4 Does the behaviour within and between clusters relate to the programme architecture?

As discussed, the engagement and performance levels in phonological awareness and early decoding games were similar, as were engagement and performance levels in games for vocabulary and comprehension processes. The basic code-related games appear at the same time in the range of options available, as do the meaning-related games in order to support the development in higher-level reading skills. These differences confirm that the underlying structure or architecture of the game is working. Students with lower DIBELS scores often demonstrate weaker basic code-related skills (University of Oregon, 2018). Thus, students with lower initial ability will be allocated to the basic skills at the start of the school year. Students with weak initial reading ability made more attempts and spent longer trying to master code-related reading skills, leading to higher levels of engagement with these particular skills. The lower levels of participation in the meaning-related skills can be attributed to these students devoting more time to developing core word reading skills, with less opportunity to engage in the meaning-related skills within the same academic year. We note that students did not play one pair of skills to the exclusion of another. This finding aligns with the understanding that reading encompasses interconnected and interdependent skills that should not be practised in isolation (Perfetti & Adlof, 2012).

5.3 Strengths, limitations and future research

Our findings demonstrate the potential to use large and complex log data to provide insights into a rich but diverse set of user's profiles and behaviours, illuminating meaningful patterns in the development of early reading skills. Previous studies have not identified a consistent single indicator of learning processes to predict learning outcomes. We have shown that it is possible to use a variety (in this case 6) engagement indicators, generated by utilizing the log data, to quantify users' behaviours, providing one way to capture and classify engagement in learning. Future research might usefully consider additional measures to capture what happens in the classrooms or at home, to determine if and how these influence engagement with digital supplements. A strength was the balanced proportion for reported male and female gender in our sample. The size of the cohort also allowed exploration of users with diverse abilities and socio-demographics. However, we note the limitation of a focus on Grade 2 students based in specific regions of the United States. Other work is needed to test the generalisability of our methods to different age groups, geographical regions, and digital apps (or similar big datasets).

We have demonstrated that unsupervised learning methods such as clustering can identify meaningful groups with varied reading outcomes. By categorizing distinct reader profiles from unstructured user log files, our study offers a methodological framework that can be adapted to other online education platforms and other apps with online log capabilities. However, we note the limitation that the clustering result can be highly influenced by types of input variables (Xu & Tian, 2015). Moreover, given the speed of development in statistical clustering there may be other clustering methods we did not consider here, which might help us to produce new ways of devising cohorts. These include approaches designed to scale to large databased, including density-based clustering algorithm (DBSCAN) (Du et al., 1999) and a combination algorithm of partitioning around medoids and clustering large applications - clustering large applications based upon randomized search (CLARANS) (Ng & Han, 2002).

Our method highlights the importance for reading development of developing a range of component skills together, a concept that matches other areas of learning. Therefore, the findings presented here can also feed into research around educational games testing logical and computation related skills in areas such as mathematics or physics facilitating further research in learning using big data more broadly (Outhwaite et al., 2023). Accordingly, our approach holds potential applicability in any discipline necessitating the combination of diverse skill sets.

Finally, we note that we grouped the code-related and meaning-related skills and considered these separately. Research has demonstrated the inter-relations between the code- and meaning-related skills examined here, for example vocabulary is related to both word reading and listening/reading comprehension (LARRC, 2015). A potential future direction of our research is to explore relationships among the skills more comprehensively.

6 Conclusion

This is the first study to use a large-scale and representative dataset to examine how user engagement is related to the development of both code-related and meaning-related reading skills in beginner readers, in a digital environment. Our findings indicate that, through active engagement in digital environments, children can develop both skill sets. The greater benefits associated with greater engagement offer insights into how to close the attainment gap amongst diverse readers through a supplementary digital reading app. Our approach demonstrates an effective analytic framework that we recommend other researchers apply to the big data that comes from the ever-increasing usage of digital reading supplements, to advance both theory and practice. Additionally, the framework under which the clustering results were organised by initial literacy ability is also recommended for future research to aid interpretability. Our data analysis, conducted independently from our industry partners, provided corroboration of both the architecture of the programme, and the benefits of repeated practice. Finally, this paper presents evidence of an effective collaboration between researchers and industry, within open science principles, demonstrating the meaningful contributions that can be made through such partnerships.

Acknowledgements

The authors acknowledge the support of the Economic and Physical Research Council who funded the first author through a PhD studentship, and Boost Amplify for supplying the dataset for analysis.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, Massachusetts: MIT Press.

Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior*, *31*(1), 542-550. https://doi.org/10.1016/j.chb.2013.05.031.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: research reviews* (pp. 77-117). Newark, Delaware: International Reading Association.

Arora, P., Virmani, D., & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, *78*, 507-512. https://doi.org/10.1016/J.PROCS.2016.02.095.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004, August/September). *Detecting student misuse of intelligent tutoring systems*. Paper presentation at the conference of the 7th Intelligent Tutoring Systems (pp. 531-540), Montreal, Canada. https://doi.org/10.1007/978-3-540-30139-4_50.

Baker, R. S. J. D., Corbett, A. T., & Aleven, V. (2008, June). *More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing*. Paper presentation at the conference of the 9th Intelligent Tutoring

Systems (pp. 406-415), Montreal, Canada. https://doi.org/10.1007/978-3-540-69132-7_44.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803. https://doi.org/10.2307/2532201.

Benton, L., Joye, N., Sumner, E., Gauthier, A., Ibrahim, S. and Vasalou, A. (2023). Exploring how children with reading difficulties respond to instructional supports in literacy games and the role of prior knowledge. *British Journal of Educational Technology*, *54*(5), 1314-1331. https://doi.org/10.1111/bjet.13318.

Biancarosa, G., & Griffiths, G. G. (2012). Technology tools to support reading in the digital age. *The Future of Children*, *22*(2), 139-160. https://doi.org/10.1353/foc.2012.0014.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345-370. https://doi.org/10.1007/BF02294361.

Ciampa, K. (2012). ICANREAD: The effects of an online reading program on Grade 1 students' engagement and comprehension strategy use. *Journal of Research on Technology in Education*, *45*(1), 27-59. https://doi.org/10.1080/15391523.2012.10782596.

de Brey, C., Musu, L., McFarland, J., Wilkinson-Flicker, S., Diliberti, M., Zhang, A., Branstetter, C., and Wang, X. (2019). *Status and trends in the education of racial and ethnic groups 2018* (NCES 2019-038). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2019/2019038.pdf. Accessed September 9, 2023

Department for Education. (2019). *National curriculum assessments: Key stage 2, 2019 (revised)*. Retrieved from https://www.gov.uk/government/statistics/national-curriculum-assessments-key-stage-2-2019-revised. Accessed September 6, 2023

Diprossimo, L., Ushakova, A., Zoski, J., Gamble, H., Irey, R., & Cain, K. (2023). The associations between child and item characteristics, use of vocabulary scaffolds, and reading comprehension in a digital environment: Insights from a big data approach. *Contemporary Educational Psychology*, *73*, 102165. https://doi.org/10.1016/J.CEDPSYCH.2023.102165.

Du, Q., Faber, V., & Gunzburger, M. (1999). Centroidal Voronoi tessellations: Applications and algorithms. *Society for Industrial and Applied Mathematics Review*, *41*(4), 637-676. https://doi.org/10.1137/S0036144599352836.

Du, X., Yang, J., Shelton, B. E., Hung, J.-L., & Zhang, M. (2021). A systematic meta-review and analysis of learning analytics research. *Behaviour & Information Technology*, *40*(1), 49-62. https://doi.org/10.1080/0144929X.2019.1669712.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363-406. https://doi.org/10.1037/0033-295X.100.3.363.

Fitzgerald, J. (1995). English-as-a-second-language learners' cognitive reading processes: A review of research in the United States. *Review of Educational Research, 65*(2), 145-190. https://doi.org/10.3102/00346543065002145.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, *41*(8), 586-588. https://doi.org/10.1093/comjnl/41.8.578.

Fraley, C., & Raftery, A. E. (2006). *MCLUST version 3: an R package for normal mixture modeling and model-based clustering*. Retrieved from https://apps.dtic.mil/sti/pdfs/ADA456562.pdf. Accessed January 3, 2023

Gómez-Aguilar, D. A., Hernández-García, Á., García-Peñalvo, F. J., & Therón, R. (2015). Tap into visual analysis of customization of grouping of activities in eLearning. *Computers in Human Behavior*, *47*, 60-67. https://doi.org/10.1016/j.chb.2014.11.001.

Goswami, U., & Bryant, P. (1990). *Phonological skills and learning to read* (1st ed.). London: Psychology Press.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading and reading disability. *Remedial and Special Education*, *7*(1), 6-10. https://doi.org/10.1177/074193258600700104.

Guterres, A. (2020). *The future of education is here*. United Nations. Retrieved from https://www.un.org/en/coronavirus/future-education-here. Accessed September 29, 2023

Halkidi, M., Vazirgiannis, M., & Hennig, C. (2015). Method-independent indices for cluster validation and estimating the number of clusters. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 616-639). New York: Chapman and Hall/CRC. https://doi.org/10.1201/b19706.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer. https://doi.org/10.1007/978-0-387-84858-7.

Hennig, C. (2023). *Package 'fpc'*. Retrieved from https://cran.r-project.org/package=fpc. Accessed May 2, 2023

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, *90*, 36-53. https://doi.org/10.1016/J.COMPEDU.2015.09.005.

Hofmann, V. (2021). App-based learning in phonological awareness and word-reading comprehension and its specific benefits for lower achieving students. International *Journal of Educational Research Open*, *2*, 100066. https://doi.org/10.1016/j.ijedro.2021.100066.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127-160. https://doi.org/10.1007/BF00401799.

Howard, S. K., Yang, J., Ma, J., Maton, K., & Rennie, E. (2018). App clusters: Exploring patterns of multiple app use in primary learning contexts. *Computers & Education*, *127*, 154-164. https://doi.org/10.1016/j.compedu.2018.08.021.

Isbister, K., & Schaffer, N. (2008). *Game usability: Advice from the experts for advancing the player experience* (1st ed.). Boca Raton: CRC Press. https://doi.org/10.1201/b14580.

Kassambara, A., & Mundt, F. (2016). *Package 'factoextra'*. Retrieved from https://piyanit.nl/wp-content/uploads/2020/10/factoextra.pdf. Accessed January 6, 2023

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning around medoids (program PAM). In L. Kaufman, & P. J. Rousseeuw (Eds.), *Finding groups in data: An introduction to cluster analysis* (pp. 68-125). Hoboken: John Wiley & Sons Inc. https://doi.org/10.1002/9780470316801.ch2.

Kendeou, P., Bohn-Gettler, C., White, M. J., & Van Den Broek, P. (2008). Children's inference generation across different media. *Journal of Research in Reading*, *31*(3), 259-272. https://doi.org/10.1111/J.1467-9817.2008.00370.X.

Kuh, G. D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New Directions for Institutional Research*, *2009*(141), 5-20. https://doi.org/10.1002/IR.283.

Kvapil, L. A., Kimpel, M. W., Jayasekare, R. R., & Shelton, K. (2022). Using Gaussian mixture model clustering to explore morphology and standardized production of ceramic vessels: A case study of pottery from Late Bronze Age Greece. *Journal of Archaeological Science: Reports*, *45*, 103543. https://doi.org/10.1016/J.JASREP.2022.103543.

Language and Reading Research Consortium (LARRC). (2015). Learning to read: Should we keep things simple? *Reading Research Quarterly*, *50*(2), 151-169. https://doi.org/10.1002/rrq.99.

Language and Reading Research Consortium (LARRC), & Logan, J. (2017). Pressure points in reading comprehension: A quantile multiple regression analysis. *Journal of Educational Psychology*, *109*(4), 451-464. https://doi.org/10.1037/edu0000150.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1-26. https://doi.org/10.1093/APPLIN/22.1.1.

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology*, *97*(2), 184-196. https://doi.org/10.1037/0022-0663.97.2.184.

Liang, D., Jia, J., Wu, X., Miao, J., & Wang, A. (2014). Analysis of learners' behaviors and learning outcomes in a massive open online course. *Knowledge Management & E-Learning: An International Journal*, *6*(3), 281-298. https://doi.org/10.34105/J.KMEL.2014.06.019.

Liao, C. H., & Wu, J. Y. (2023). Learning analytics on video-viewing engagement in a flipped statistics course: Relating external video-viewing patterns to internal motivational dynamics and performance. *Computers & Education*, *197*, 104754. https://doi.org/10.1016/J.COMPEDU.2023.104754.

Lim, H., Kim, S., Chung, K. M., Lee, K., Kim, T., & Heo, J. (2022). Is college students' trajectory associated with academic performance? *Computers & Education*, *178*, 104397. https://doi.org/10.1016/J.COMPEDU.2021.104397.

Locher, F., & Pfost, M. (2020). The relation between time spent reading and reading comprehension throughout the life course. *Journal of Research in Reading*, *43*(1), 57-77. https://doi.org/10.1111/1467-9817.12289.

López-Escribano, C., Valverde-Montesino, S. and García-Ortega, V. (2021). The impact of e-book reading on young children's emergent literacy skills: An analytical review. *International Journal of Environmental Research and Public Health*, *18*(12), 6510. https://doi.org/10.3390/ijerph18126510.

MacQueen, J. (1967, January). *Some methods for classification and analysis of multivariate*. Paper presentation at the Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability: Volume 1 (pp. 281-297). Oakland, CA, USA. http://projecteuclid.org/euclid.bsmsp/1200512992.

Melby-Lervåg, M., & Lervåg, A. (2014). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, *140*(2), 409-433. https://doi.org/10.1037/A0033890.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*(2), 181-204. https://doi.org/10.1007/BF01897163.

Moubayed, A., Injadat, M., Shami, A., & Lutfiyya, H. (2020). Student engagement level in an e-Learning environment: Clustering using k-means*. American Journal of Distance Education*, *34*(2), 137-156. https://doi.org/10.1080/08923647.2020.1696140.

Muter, V., Hulme, C., Snowling, M., & Taylor, S. (1998). Segmentation, not rhyming, predicts early progress in learning to read. *Journal of Experimental Child Psychology*, *71*(1), 3-27. https://doi.org/10.1006/JECP.1998.2453.

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common core state standards (English Language Arts Standards)*. Retrieved from http://corestandards.org/. Accessed September 9, 2023

National Reading Panel (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. National Institute of Child Health and Human Development, National Institutes of Health. Retrieved from https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/Documents/report.pdf. Accessed June 6, 2023

Newton, S., Gamble, H., Su, Y., Zoski, J., & Damico, D. (2019). *Examining the Impact of Amplify Reading on Student Literacy in Grades K-2. 2019 Report* (ED604917). ERIC. Retrieved from https://files.eric.ed.gov/fulltext/ED604917.pdf. Accessed December 9, 2022

Ng, R. T., & Han, J.(2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, *14*(5), 1003-1016. https://doi.org/10.1109/TKDE.2002.1033770.

Nizam, D. N. M., & Law, E. L. C. (2021). Derivation of young children's interaction strategies with digital educational games from gaze sequences analysis. *International Journal of Human-Computer Studies*, *146*, 102558. https://doi.org/10.1016/j.ijhcs.2020.102558.

Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading*, *16*(2), 91-121. https://doi.org/10.1080/10888438.2010.529219.

Outhwaite, L. A., Early, E., Herodotou, C., & Van Herwegen, J. (2023). Understanding how educational maths apps can enhance learning: A content analysis and qualitative comparative analysis. *British Educational Research Association Journal*, *54*(5), 1292-1313. https://doi.org/10.1111/bjet.13339.

Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336-3341. https://doi.org/10.1016/J.ESWA.2008.01.039.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227-247). Oxford: Blackwell. https://doi.org/10.1002/9780470757642.CH13.

Perfetti, C., & Adlof, S. M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. In J. Sabatini, E. Albro, & T. O'Reilly (Ed.), *Measuring up: Advances in how we assess reading ability* (pp. 3-20). Lanham, Maryland: Rowman & Littlefield Education.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/index.html. Accessed September 3, 2022

Sánchez, J. L. G., Vela, F. L. G., Simarro, F. M., & Padilla-Zea, N. (2012). Playability: Analysing user experience in video games. *Behaviour & Information Technology*, *31*(10), 1033-1054. https://doi.org/10.1080/0144929X.2012.710648.

Saqr, M., & López-Pernas, S. (2021). The longitudinal trajectories of online engagement over a full program. *Computers & Education*, *175*, 104325. https://doi.org/10.1016/J.COMPEDU.2021.104325.

Schlechty, P. C. (2001). *Shaking up the school house: How to support and sustain educational innovation*. San Francisco, CA: Jossey-Bass Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464. https://doi.org/10.1214/aos/1176344136.

Shirkhorshidi, A. S., Aghabozorgi, S., & Ying Wah, T. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLOS ONE*, *10*(12), e0144059. https://doi.org/10.1371/JOURNAL.PONE.0144059.

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist*, *50*(1), 1-13. https://doi.org/10.1080/00461520.2014.1002924.

Snow, C. E., Burns, M. S., & Grifin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: The National Academies Press. https://doi.org/10.17226/6023.

Sparks, R. L., Patton, J., & Ganschow, L. (2012). Profiles of more and less successful L2 learners: A cluster analysis study. *Learning and Individual Differences*, *22*(4), 463-472. https://doi.org/10.1016/J.LINDIF.2012.03.009.

Šulc, Z., & Řezanková, H. (2019). Comparison of similarity measures for categorical data in hierarchical clustering. *Journal of Classification*, *36*(1), 58-72. https://doi.org/10.1007/s00357-019-09317-5.

The No Child Left Behind Act of 2001. (2002). *For no child left behind act: the No Child Left Behind Act's Reading First Intiative*. Public Law No. 107-110, 115 Stat. 1425. Retrieved from https://www.govinfo.gov/content/pkg/PLAW-107publ110/pdf/PLAW-107publ110.pdf. Accessed May 5, 2023

Theodoridis, S., & Koutroumbas, K. (2006). *Pattern recognition* (3rd ed.). Elsevier Science & Technology. https://doi.org/10.1016/B978-0-12-369531-4.X5000-8.

University of Oregon. (2018). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS)*. Center on Teaching and Learning. Eugene, Oregon: University of Oregon. https://dibels.uoregon.edu.

Vardal, O., Bonometti, V., Drachen, A., Wade, A., & Stafford, T. (2022). Mind the gap: Distributed practice enhances performance in a MOBA game. *PLOS ONE*, *17*(10), e0275843. https://doi.org/10.1371/JOURNAL.PONE.0275843.

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading*, *15*(1), 8-25. https://doi.org/10.1080/10888438.2011.536125.

Vnucko, G., & Klimova, B. (2023). Exploring the potential of digital game-based vocabulary learning: A systematic review. *Systems*, *11*(2), 57. https://doi.org/10.3390/systems11020057.

Wang, X., Liu, Q., Pang, H., Tan, S. C., Lei, J., Wallace, M. P., & Li, L. (2023). What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis. *Computers & Education*, *194*, 104703. https://doi.org/10.1016/J.COMPEDU.2022.104703.

Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L. D. A., François, R., et al. (2019a). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/JOSS.01686.

Wickham, H., François, R., Henry, L., & Müller, K. (2019b). *Package 'dplyr'*. Retrieved from https://cran.r-hub.io/web/packages/dplyr/dplyr.pdf. Accessed January 3, 2023

Woolfolk, A. (2007). *Educational psychology* (10th ed.). Boston: Allyn and Bacon.

Wyse, D., Jones, R., Bradford, H., & Wolpert, M. A. (2013). *Teaching English, language and literacy* (3rd ed.). London: Routledge. https://doi.org/10.4324/9780203073520.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*, 165-193. https://doi.org/10.1007/S40745-015-0040-1.

Yang, D., Zargar, E., Adams, A. M., Day, S. L., & Connor, C. M. D. (2020). Using interactive e-book user log variables to track reading processes and predict digital learning outcomes. *Assessment for Effective Intervention*, *46*(4), 292-303. https://doi.org/10.1177/1534508420941935.

Yoo, C., Yoo, Y., Um, H. Y., & Yoo, J. K. (2020). On hierarchical clustering in sufficient dimension reduction. *Communications for Statistical Applications and Methods*, *27*(4), 431-443. https://doi.org/10.29220/CSAM.2020.27.4.431.

Zheng, B., Lin, C. H., & Kwon, J. B. (2020). The impact of learner-, instructor-, and course-level factors on online learning. *Computers & Education*, *150*, 103851. https://doi.org/10.1016/J.COMPEDU.2020.103851.

Appendices

Variables	Description
Engagement Variables	
All hours spent on skills	Sum of rounds duration of playing across skills (conversion of units to hours).
Variety of games played within skills	The number of types of games are played within a specific skill.
All attempts within skills	Count the number of reattempting rounds for different games and sum the values within a skill for each user.
The number of levels played within skills	Count the number of different levels has been played for different games and sum the values within a skill for each user.
Days of playing within app	Count the number of days the user engaged in games within 4 skills.
Proportion of early exit rounds	Take ratio of the number of the early exits' rounds over *all attempts within skills.*
Performance Variables	
Time to mastery	The time from the first attempt to first mastery.
Number of levels mastered	Sum the total number of levels mastered within skill games.
Speed to mastery	Represents the average number of levels a group of students can master within one hour. Take ratio of *number of levels mastered* over *all hours spent within skills*.

Attempts to mastery	Denote the attempts needed to master a level. Take ratio of *all attempts* over *number of levels mastered*.

Table A.1. The description for each created variable.

		Well below benchmark	Below benchmark	At benchmark	Above benchmark	Total
Sample size (%)		7657 (39.66%)	3260 (16.89%)	5206 (26.97%)	3182 (16.48%)	19305
Number of male (M) and female (F) (%)	M (%)	3529 (46.09%)	1485 (45.55%)	2436 (46.79%)	1561 (49.06%)	9011 (46.68%)
	F (%)	3450 (45.06%)	1582 (48.53%)	2484 (47.71%)	1471 (46.23%)	8987 (46.55%)
Number of English language learner and (ELL) non-English language learner (non-ELL) (%)	ELL (%)	2762 (36.07%)	944 (28.96%)	920 (17.67%)	287 (9.02%)	4913 (25.45%)
	non-ELL (%)	3695 (48.26%)	1931 (59.20%)	3717 (71.40%)	2635 (82.81%)	11978 (62.05%)
Number of special education needs (SEN) and non-special	SEN (%)	852 (11.13%)	246 (7.55%)	295 (5.67%)	118 (3.71%)	1511 (7.83%)
	non-SEN (%)	5685 (74.25%)	2657 (81.50%)	4379 (84.11%)	2833 (89.03%)	15554 (80.57%)

education needs (non-SEN) (%)						
Proportion of races (%) American Indian (AI) Alaskan Native (AN) Asian (AS) Black or African American (B) Hispanic or Latino (H) Native Hawaiian or Pacific Islander (NHPI) Multiracial/other (M) White (W) Not Specified (NS)	AI (%)	14 (0.18%)	2 (0.06%)	7 (0.13%)	7 (0.22%)	30 (0.16%)
	AN (%)	2 (0.03%)	0 (0%)	2 (0.04%)	1 (0.03%)	5 (0.03%)
	AS (%)	127 (1.66%)	88 (2.70%)	212 (4.07%)	219 (6.88%)	646 (3.35%)
	B (%)	824 (10.76%)	283 (8.68%)	437 (8.39%)	252 (7.92%)	1796 (9.30%)
	H (%)	4962 (64.80%)	2192 (67.24%)	3392 (65.16%)	1885 (59.24%)	12431 (64.39%)
	NHPI (%)	48 (0.63%)	46 (1.41%)	133 (2.55%)	160 (5.03%)	387 (2.00%)
	M (%)	128 (1.67%)	44 (1.35%)	43 (0.83%)	25 (0.79%)	240 (1.24%)
	W (%)	553 (7.22%)	269 (8.25%)	456 (8.76%)	375 (11.79%)	1653 (8.56%)
	NS (%)	8 (0.10%)	2 (0.06%)	4 (0.08%)	4 (0.13%)	18 (0.09%)

Table A.2. A summary of the social demographics of the sample (Grade 2 students). Missing records are not included.

Skills	*Ranking of all hours spent (v1)*	*Ranking of variety of games played (v2)*	*Ranking of all attempts (v3)*	*Ranking of number of levels played (v4)*	*Ranking of days of playing (v5)*	*Ranking of proportion of early exit rounds (v6)*	*Engagement ranking in radar charts (see Figure 5&6) (average ranking of v1-v5)*
Cluster 1							
Phonological awareness	7	8	7	7	7	3	7
Early decoding	8	8	7	7	7	2	7
Vocabulary	3	3	3	3	3	7	3
Comprehension processes	3	3	3	3	3	7	3
Cluster 2							
Phonological awareness	5	2	4	4	4	5	4
Early decoding	5	5	5	4	5	4	5
Vocabulary	1	1	2	1	2	9	1
Comprehension processes	1	1	1	1	1	8	1

Cluster 3							
Phonological awareness	8	9	8	8	8	2	8
Early decoding	7	7	6	6	6	3	6
Vocabulary	6	5	6	5	6	4	6
Comprehension processes	6	5	6	6	6	6	6
Cluster 4							
Phonological awareness	9	7	9	9	9	1	9
Early decoding	9	9	9	9	9	1	9
Vocabulary	8	8	8	8	8	2	8
Comprehension processes	9	9	9	9	9	1	9
Cluster 5							
Phonological awareness	2	1	2	1	2	9	2
Early decoding	1	1	1	1	1	9	1
Vocabulary	2	2	1	2	1	8	2
Comprehension processes	2	2	2	2	2	9	2

Cluster 6							
Phonological awareness	4	6	5	6	6	4	6
Early decoding	3	2	3	2	3	6	3
Vocabulary	4	4	4	4	4	6	4
Comprehension processes	4	4	4	4	4	5	4
Cluster 7							
Phonological awareness	6	5	6	5	5	6	5
Early decoding	6	6	8	8	8	8	8
Vocabulary	9	9	9	9	9	1	9
Comprehension processes	8	8	8	8	8	2	8
Cluster 8							
Phonological awareness	3	4	3	3	3	7	3
Early decoding	4	4	4	5	4	5	4
Vocabulary	7	7	7	7	7	3	7
Comprehension processes	7	7	7	7	7	3	7

Cluster 9

Phonological awareness	1	3	1	2	1	8	1
Early decoding	2	3	2	3	2	7	2
Vocabulary	5	6	5	6	5	5	5
Comprehension processes	5	6	5	5	5	4	5

Table A.3. Clustering results of the ranking of engagement for each cluster across four skills (k-means).

Supplementary materials

	Variable Description
User id	A unique identifier for a student.
Skill	A description of the type of content the games aim to teach. One of phonological awareness, early decoding, vocabulary, and comprehension processes.
Game id	A unique identifier for the game in which a reveal word appears.
Game level	An integer identifying the level of a game in within the student responded to the question.
Is level mastered	True if the student's score (proportion of correct) exceeded the mastery threshold for the game level in question. This threshold varies by game but is usually .8, which is the criterion built into each game by the developers. Otherwise, false.
Items answered	A count of the number of questions a student responded to.
Items correct	A count of the number of correct question responses the student submitted.
Score	(Proportion of correct) Take ratio of items correct to items answered.
Round started at	(Timestamps) The Universal Time Coordinated datetime when the student initiated the round.

Round ended at	(Timestamps) The Universal Time Coordinated datetime when the student completed the round.
Elapsed time min	The number of minutes that elapsed between the start and end of the round. The elapsed time is calculated by summing the amount of time between data collection moments in the app; if any of these interstitial times exceeds 20 minutes, they are set to 0.
Is early exit	TRUE if the student exited the round without submitting a response to every item in the level. Otherwise, FALSE.
Mind the Gap performance level	The ordering of the quintile ranking of the percentile rank of the difficulty adjusted score the student achieved on the passage, rounded to the nearest .5 points.
Mind the Gap start	The first result of Mind the Gap performance level for each user not necessary around September 2020.
Mind the Gap end	The final result of Mind the Gap performance level for each user around September 2021.
Mind the Gap difference	Mind the Gap end - Mind the Gap start
Student Characteristics	
Student gender	The student's gender. One of M, F, N/A or missing. M=Male, F = Female.

Is ell	1 = The student is an English as a second language student; 0 = The student is not an English as a second language students.
Is special education	1 = The student receives Special Education Needs; 0 = The student does not receive Special Education Needs.
Race	AN = Alaskan Native; AI = American Indian; AS = Asian; NHPI= Native Hawaiian or Pacific Islander; B= Black or African American; H = Hispanic or Latino; W = White; M = Multiracial/other; N/A= Not Available; NS = Not Specified
DIBELS composite performance level (initial assessment)	The performance level of the composite score from the DIBELS result. One of Well Below Benchmark < Below Benchmark < At Benchmark < Above Benchmark

Table S.1. Log file variables.

Engagement Variable	**Range**	**Median**	**Mean**	**SD**
All hours spent on phonological awareness	(0, 20.20)	0.16	0.42	0.65
All hours spent on early decoding	(0, 22.32)	0.53	1.21	1.69
All hours spent on vocabulary	(0, 16.53)	0.20	0.69	1.24
All hours spent on comprehension processes	(0, 29.47)	0.41	0.97	1.49
Variety of games played within phonological awareness	(0, 5.00)	1.00	0.91	0.93
Variety of games played within early decoding	(0, 8.00)	2.00	2.80	2.21
Variety of games played within vocabulary	(0, 6.00)	1.00	1.72	1.65
Variety of games played within comprehension processes	(0, 11.00)	2.00	2.52	2.36
All attempts within phonological awareness	(0, 614.00)	4.00	9.74	16.16
All attempts within early decoding	(0, 651.00)	11.00	26.54	39.27
All attempts within vocabulary	(0, 514.00)	7.00	20.76	35.75
All attempts within comprehension processes	(0, 883.00)	9.00	22.75	37.77
The number of levels played within phonological awareness	(0, 70.00)	2.00	2.91	3.63

The number of levels played within early decoding	(0, 160.00)	6.00	9.76	11.61
The number of levels played within vocabulary	(0, 94.00)	3.00	7.81	11.00
The number of levels played within comprehension processes	(0, 74.00)	5.00	7.53	9.33
Days of playing within phonological awareness	(0, 56.00)	2.00	3.59	4.91
Days of playing within early decoding	(0, 86.00)	5.00	9.12	10.73
Days of playing within vocabulary	(0, 90.00)	3.00	7.46	10.26
Days of playing within comprehension processes	(0, 98.00)	4.00	6.99	8.95
Proportion of early exit rounds within phonological awareness (%)	(0%, 100%)	0%	15.86%	0.22
Proportion of early exit rounds within early decoding (%)	(0%, 100%)	16.91%	22.32%	0.22
Proportion of early exit rounds within vocabulary (%)	(0%, 100%)	15.15%	21.64%	0.24
Proportion of early exit rounds within comprehension processes (%)	(0%, 100%)	14.29%	22.80%	0.26

Table S.2. Summary statistics for the six engagement variables obtained each skill.

Cluster algorithms	Size	Average silhouette width
Gaussian mixture models with 9 clusters	2153	-0.14
	2508	-0.001
	2521	-0.09
	3841	-0.08
	1238	-0.11
	2159	-0.005
	3618	-0.06
	764	0.53
	503	0.58
K-means with 9 clusters	1726	0.19
	571	0.06
	3468	0.12
	4862	0.33
	248	0.03
	1126	0.11
	4001	0.10
	2516	0.10
	787	0.05
Partition around medoids with 9 clusters	2303	0.06
	1489	0.13
	3182	0.16
	2845	0.06
	2249	0.09

	2368	0.03
	3076	0.38
	945	0.12
	848	-0.06
Clustering large applications with 9 clusters	3900	0.03
	4972	0.05
	3319	0.03
	431	0.17
	1289	0.16
	723	0.12
	3423	0.28
	42	0.04
	1206	-0.10
Hierarchical clustering with 9 clusters	19277	0.68
	13	0.14
	1	0.00
	1	0.00
	5	0.09
	4	0.24
	2	0.58
	1	0.00
	1	0.00

Table S.3. Cohorts size and average silhouette width.

		Cluster 1 (n = 1726)	Cluster 2 (n = 571)	Cluster 3 (n = 3468)	Cluster 4 (n = 4862)	Cluster 5 (n = 248)	Cluster 6 (n = 1126)	Cluster 7 (n = 4001)	Cluster 8 (n = 2516)	Cluster 9 (n = 787)
Number of male (M) and female (F)	M (%)	48.96	54.47	49.28	46.24	47.58	47.42	43.74	42.45	49.05
	F (%)	46.06	41.86	46.14	45.99	44.35	44.14	48.16	50.48	45.11
Number of English language learner and (ELL) non-English language learner (non-ELL)	ELL (%)	10.72	18.04	15.46	23.51	39.92	25.75	35.99	32.39	38.37
	Non-ELL (%)	73.24	70.93	72.35	65.55	47.18	55.86	53.44	53.50	48.67
Number of special	SEN (%)	4.69	6.48	4.64	6.09	10.89	8.53	11.00	10.37	14.23

education needs (SEN) and non-special education needs (non-SEN)	Non-SEN (%)	81.23	83.01	84.34	83.34	76.61	73.98	79.55	76.95	73.57
Proportion of races: American Indian (AI) Alaskan Native (AN) Asian (AS) Black or African American (B) Hispanic or Latino (H)	AI (%)	0.29	0	0	0.10	0.40	0.09	0.17	0.24	0.25
	AN (%)	0	0	0.06	0.04	0	0	0	0	0.13
	AS (%)	7.99	8.76	4.38	2.63	4.03	2.58	1.75	2.03	2.29
	B (%)	7.65	6.48	8.25	9.05	5.24	8.26	10.02	11.84	12.20
	H (%)	55.21	59.37	61.51	65.47	67.74	62.70	69.88	65.10	65.44

Native Hawaiian or Pacific Islander (NHPI) Multiracial/other (M) White (W) Not Specified (NS)	NHPI (%)	4.17	3.15	3.26	2.16	0.81	1.24	0.95	0.79	0.64
	M (%)	1.22	0.70	1.56	0.64	2.02	1.87	1.00	1.47	3.43
	W (%)	12.00	13.31	9.69	8.31	9.68	10.12	5.82	7.79	8.01
	NS (%)	0.12	0	0.12	0.02	0	0.27	0.10	0.16	0

Table S.4. Cluster’s characteristics.

Variables	Description
Engagement Variables	
All hours spent	From the longest hours spent to the shortest hours spent on one skill.
Variety of games played	From the most variety of games to the least variety of games played within one skill.
All attempts	From the most attempts to the least attempts within one skill.
The number of levels played	From the greatest number of levels played to the least number of levels played within one skill.
Days of playing	From the greatest days of playing to the least days dedicated to playing within one skill.
Proportion of early exit rounds	From the least proportion of early exit rounds to the most proportion of early exit rounds.
Performance Variables	
Number of levels mastered	From the most levels mastered to the least levels mastered.
Time to mastery	From the shortest time required to master a level to the longest time required to master a level.
Attempts to mastery	From the least attempts required to master a level to the most attempts required to master a level.
Speed to mastery	From the most levels mastered per hour to the least levels mastered per hour.

Table S.5. Ranking rule for both engagement indicators and performance indicators from the best to the worst (1st to 9th).

Skills	*All hours spent*	*Variety of games played*	*All attempts*	*The number of levels played*	*Days of playing*
Cluster 1					
Phonological awareness	0.13	0.44	3.12	1.26	1.16
Early decoding	0.57	1.81	13.25	7.13	5.39
Vocabulary	2.12	4.35	56.66	23.29	21.08
Comprehension processes	2.72	6.22	63.23	22.83	19.89
Cluster 2					
Phonological awareness	0.44	1.94	14.65	5.72	5.26
Early decoding	1.15	3.62	30.29	14.55	10.53
Vocabulary	4.70	5.40	133.48	41.52	36.40
Comprehension processes	5.47	8.45	134.89	37.03	33.26
Cluster 3					
Phonological awareness	0.08	0.31	1.86	0.79	0.80
Early decoding	0.64	1.99	14.83	7.43	5.79
Vocabulary	0.55	2.52	17.23	8.35	7.51
Comprehension processes	0.93	3.36	20.99	8.25	7.64

Cluster 4					
Phonological awareness	0.07	0.44	1.40	0.72	0.74
Early decoding	0.25	1.26	5.19	2.83	2.35
Vocabulary	0.11	0.72	3.34	1.66	1.60
Comprehension processes	0.10	0.60	1.82	1.04	0.89
Cluster 5					
Phonological awareness	1.54	2.74	49.63	12.64	12.64
Early decoding	6.72	7.58	181.58	53.69	43.07
Vocabulary	4.62	4.59	155.97	38.75	39.27
Comprehension processes	5.04	7.09	134.65	25.33	28.19
Cluster 6					
Phonological awareness	0.46	0.73	10.44	3.25	3.38
Early decoding	4.10	7.05	101.57	36.57	29.95
Vocabulary	1.55	2.80	48.98	17.24	18.33
Comprehension processes	1.65	3.65	38.91	11.48	10.70
Cluster 7					
Phonological awareness	0.42	1.43	9.41	3.47	3.84

Early decoding	0.71	2.24	13.01	4.56	5.30
Vocabulary	0.05	0.32	1.77	0.85	0.75
Comprehension processes	0.17	1.04	3.58	1.79	1.67
Cluster 8					
Phonological awareness	1.08	1.52	23.68	6.46	9.03
Early decoding	2.10	4.84	41.34	14.50	15.28
Vocabulary	0.31	1.32	10.85	4.35	4.49
Comprehension processes	0.86	2.45	19.90	6.97	6.67
Cluster 9					
Phonological awareness	2.23	1.87	50.38	9.59	16.32
Early decoding	4.99	6.16	101.75	24.26	31.96
Vocabulary	0.85	1.96	28.75	8.11	10.60
Comprehension processes	1.53	2.96	37.84	9.14	10.49

Table S.6. Clustering results of the engagement values for each cluster across four skills (k-means).

Skills	*Attempts to mastery*	*Number of levels mastered*	*Speed to mastery (average number of levels mastered per hour)*	*Time to mastery (min)*
Cluster 1				
Phonological awareness	1.72	1.01	7.52	4.80
Early decoding	1.40	6.35	13.31	3.94
Vocabulary	1.55	19.24	10.21	3.99
Comprehension processes	1.54	18.45	7.63	5.42
Cluster 2				
Phonological awareness	1.38	4.67	16.74	3.17
Early decoding	1.36	12.81	13.86	5.58
Vocabulary	1.61	35.32	8.59	4.49
Comprehension processes	1.75	30.62	6.36	5.88
Cluster 3				
Phonological awareness	1.99	0.63	5.39	5.64
Early decoding	1.47	6.48	12.30	4.19
Vocabulary	1.53	6.76	14.59	3.21
Comprehension processes	1.63	6.27	7.78	5.24

Cluster 4				
Phonological awareness	2.28	0.43	6.85	6.16
Early decoding	1.71	2.22	9.42	4.62
Vocabulary	1.53	1.31	9.43	3.15
Comprehension processes	1.63	0.82	5.11	4.98
Cluster 5				
Phonological awareness	1.77	10.50	9.83	4.64
Early decoding	2.01	45.94	7.89	5.66
Vocabulary	2.09	31.89	8.31	4.57
Comprehension processes	2.11	19.87	4.62	7.07
Cluster 6				
Phonological awareness	2.16	2.60	5.25	6.70
Early decoding	1.89	31.15	8.47	5.51
Vocabulary	2.00	14.10	11.14	4.22
Comprehension processes	1.82	9.16	6.88	5.94
Cluster 7				
Phonological awareness	2.45	2.15	6.73	7.07

Early decoding	2.07	3.02	5.88	7.11
Vocabulary	1.67	0.67	4.35	3.18
Comprehension processes	1.58	1.36	8.01	4.77
Cluster 8				
Phonological awareness	2.85	4.75	5.28	9.20
Early decoding	2.09	11.48	6.59	7.45
Vocabulary	1.79	3.36	9.33	3.55
Comprehension processes	1.76	5.42	8.06	5.24
Cluster 9				
Phonological awareness	3.58	7.56	3.96	12.38
Early decoding	2.80	19.32	4.63	10.35
Vocabulary	2.32	6.15	8.34	4.95
Comprehension processes	2.13	7.20	6.33	6.43

Table S.7. Clustering results of the performance values for each cluster across four skills (k-means).

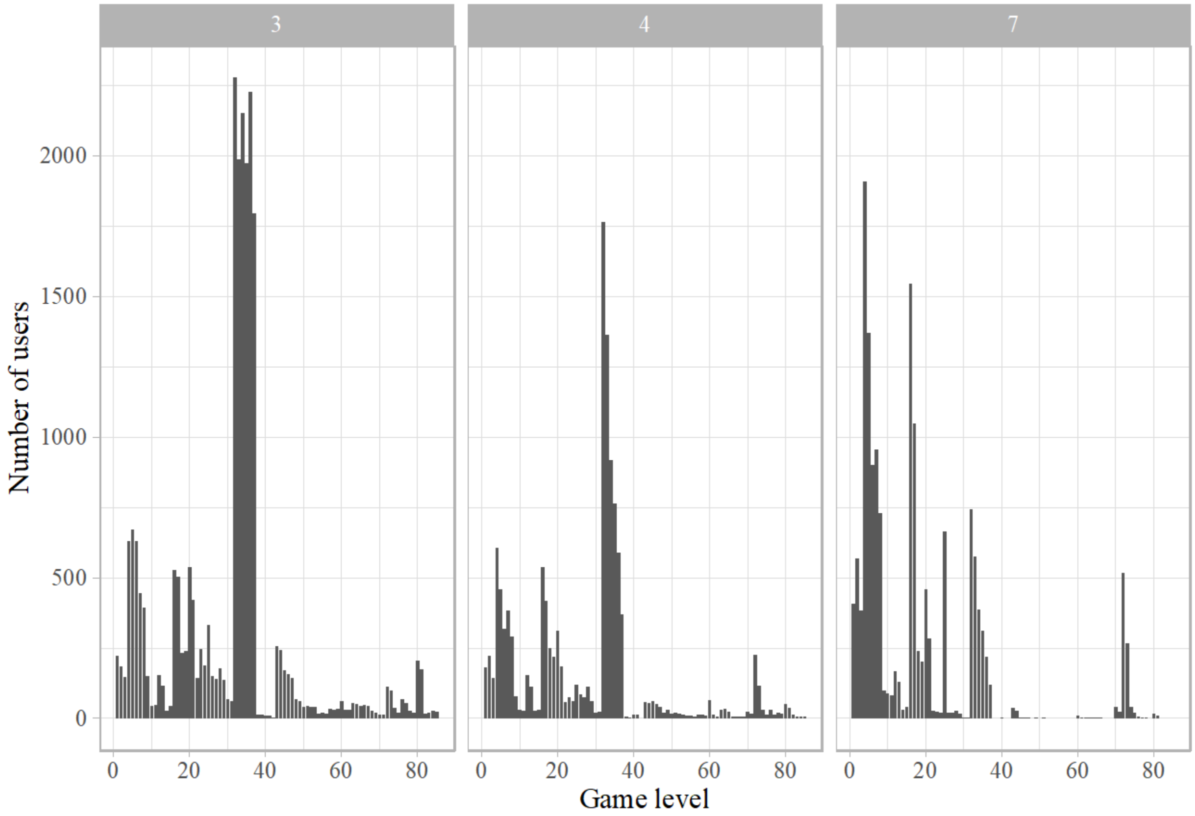


Figure S.8. The number of individuals from Clusters 3, 4, and 7 who played each level of the early decoding games.

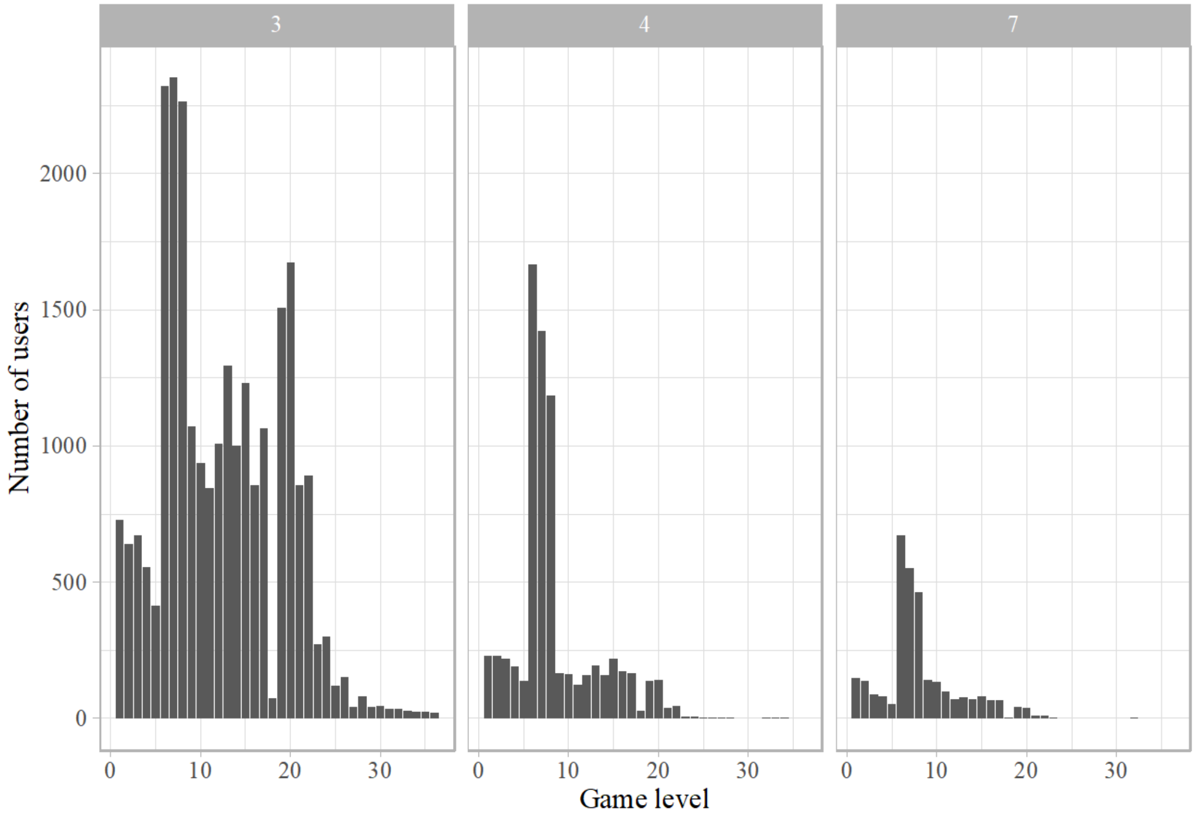


Figure S.9. The number of individuals from Cluster 3, 4, and 7 who played each level of the vocabulary games.