Article:
Foldnes, N., Uppstad, P.H., Grønneberg, S. et al. (1 more author) (2024) School entry detection of struggling readers using gameplay data and machine learning. Frontiers in Education, 9. 1487694. ISSN 2504-284X

https://doi.org/10.3389/feduc.2024.1487694

frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 22 November 2024
DOI 10.3389/feduc.2024.1487694

OPEN ACCESS

EDITED BY
Mohammad Khalil,
University of Bergen, Norway

REVIEWED BY
Quan Zhang,
Jiaxing University, China
Ioannis Dimakos,
University of Patras, Greece

*CORRESPONDENCE
Njål Foldnes
✉ njal.foldnes@gmail.com

RECEIVED 28 August 2024
ACCEPTED 23 October 2024
PUBLISHED 22 November 2024

CITATION
Foldnes N, Uppstad PH, Grønneberg S and Thomson JM (2024) School entry detection of struggling readers using gameplay data and machine learning. *Front. Educ.* 9:1487694. doi: 10.3389/feduc.2024.1487694

School entry detection of struggling readers using gameplay data and machine learning

Njål Foldnes[1]*, Per Henning Uppstad[1], Steffen Grønneberg[2] and Jenny M. Thomson[3]

[1]Norwegian Centre for Reading Education and Research, University of Stavanger, Stavanger, Norway, [2]BI Norwegian Business School, Oslo, Norway, [3]School of Allied Health Professions, Nursing and Midwifery, The University of Sheffield, Sheffield, United Kingdom

Introduction: Current methods for reading difficulty risk detection at school entry remain error-prone. We present a novel approach utilizing machine learning analysis of data from GraphoGame, a fun and pedagogical literacy app.

Methods: The app was played in class daily for 10 min by 1,676 Norwegian first graders, over a 5-week period during the first months of schooling, generating rich process data. Models were trained on the process data combined with results from the end-of-year national screening test.

Results: The best machine learning models correctly identified 75% of the students at risk for developing reading difficulties.

Discussion: The present study is among the first to investigate the potential of predicting emerging learning difficulties using machine learning on game process data.

KEYWORDS

early detection, reading, machine learning, process data, reading difficulties

1 Introduction

Learning to read is a crucial skill acquired during the first years of school and children with difficulties in acquiring this skill may face adverse educational, vocational, and health outcomes (McLaughlin et al., 2014; DeWalt et al., 2004). While researchers have pointed to the relative ease with which a majority of students learn to read (Shankweiler and Liberman, 1989), the same process is extremely effortful for struggling readers and more so for students with dyslexia. If difficulties in this learning process are not identified promptly and ameliorated, the performance gap between struggling readers and their non-struggling peers will only widen, a phenomenon coined, the "Matthew Effect" (Stanovich, 2009). Much research has therefore focused on how to optimize early and accurate identification of struggling readers. The most common form of early detection tool is currently a one-time multi-component assessment where early reading skills and their precursors are examined (Thompson et al., 2015; Phillips et al., 2009). Learning to read involves mastering a number of different component skills over a protracted time course; these component skills include letter-sound decoding, whole word recognition, reading fluency, and the ultimate goal of reading—comprehension (Scarborough et al., 2009). As the statistical analysis capacity of reading research has expanded, we are increasingly able to identify and quantify the relative contributions of these different factors at different points of the developmental process. However, while substantive progress has been made, current detection tools remain error-prone at the level of individual prediction, potentially resulting in both false positives,

where children are identified as having a difficulty which may have resolved without expert input, or false negatives, where a child who will go on to have persistent difficulties is not identified as "at-risk" by an assessment tool. Arguably, a false negative classification may have more serious consequences for the individual student than a false positive classification. Still, both these types of mis-classification can have significant socio-emotional and educational consequences.

A currently underutilized opportunity is making use of the extensive data made available via tracking children's learning within educational technology, for the purposes of detecting reading difficulty. With an increasing number of digital supports for literacy learning (Livingstone, 2021; Yang et al., 2018) being present in the classroom, new opportunities for observing and evaluating children's pathway to literacy occur—at a level of detail and specificity previously unavailable. One potential advantage of this approach is that the game log data can non-intrusively capture the dynamic process of learning, allowing for assessment of a phenomenon more closely akin to the process of learning to read itself. An additional characteristic of this type of data exploitation is that the method of assessment can go hand-in-hand with play-based learning—a process which has been called "stealth assessment" in the wider learning analytics literature (Shute et al., 2021). The processing of such complex data also necessitates more powerful statistical methods that are underutilized in current reading research; and the aim of the present study was to employ and compare multiple machine learning approaches to analyzing process data from the serious literacy app GraphoGame.

1.1 Related work

1.1.1 Current approaches to early identification

Precise early identification of reading difficulties has proven challenging for a number of reasons. Typically, early identification is wanted at the time of school entry or during the first year of school, and is typically maintained as a one-time multi-component assessment. Early identification of reading difficulty is, however, dependent on using early predictors of reading, as the identification is likely to take place before formal reading instruction has started. Research has documented a set of such pre-reading skills that predicts later reading development in alphabetical orthographies (Caravolas et al., 2013; Solheim et al., 2021). Still, we need to improve identification for this age group, as identification is likely more effective when provided early (Lovett et al., 2017). Dixon et al. (2022) emphasize that early identification before, or at the onset of, reading instruction is hampered by floor effects and are insensitive to childrens' learning experience and opportunities. One-time multi-component assessments applied typically includes phonological awareness, letter knowledge or word decoding, verbal short-time memory, rapid automatized naming and oral language. In a comparison of state of the art prediction across Norway and Finland (Solheim et al., 2021), a typical logistic predictive model identified 42% at risk at the end of first grade in Finland, but the same model predicted only 27.9% in Norway. The authors point to the need for language and context specific predictors, when/if following this test-tradition further. Contrasting the one-time multi-component assessment approach, dynamic assessment, i.e., assessment of process where feedback is given alongside, "...and may be less dependent on learning background" (Dixon et al., 2022, p. 1). In dynamic assessment, rather than assessing the products of learning, the focus is the processing of learning itself—how does an individual respond to both instruction and feedback. This approach is conceptually grounded in Vygotsky's construct of Zones of Proximal Development (Dumas et al., 2020), which can be defined as the space between what a learner can do without assistance, and what a learner can do with feedback or support. While a such construct can be measured also at a single time point, dynamic assessment emphasizes the aspect of learning over time, involving frequent feedback and multiple measure points allowing for learning to take place and be measured. Applying psychometric rigor to such an individualized assessment process is not without its challenges (Grigorenko and Sternberg, 1998), however as assessment design advances, the potential of dynamic assessment is growing. Indeed, recent meta-analysis (Dixon et al., 2022) shows that dynamic assessment may provide better prediction than the static approach, while calling for future studies to adopt longer developmental windows for the assessment.

1.1.2 Early identification utilizing dynamic literacy gameplay data

One unique opportunity to address this call—combining rich, multi-faceted process data, collected over a longer window of time—is exploring the predictive capabilities of dynamic assessment using logdata from the digital reading games. The last few decades have seen a steady increase in the availability of digital early reading games, as well as an evidence base that supports their use (Verhoeven et al., 2022). This provides unprecedented and largely untapped information on children's learning processes.

Graphogame is a digital learning platform that is singular in terms of its combination of global reach and accessibility of player logdata for research and innovation. It is a play-like, digital learning platform that provides children with training in phoneme awareness, letter-sound and early word decoding training, providing real-time feedback. It was originally devised by researchers at the University of Jyväskylä in Finland (Lyytinen et al., 2009, 2007). Since its inception in Finland and promising initial findings, the game has subsequently been adapted for at least 10 alphabetic languages of varying orthographic depth, across more than 20 countries in four continents (Africa, Europe, North America, South America).

The game content adapts to the individual player according to actual performance in identifying letters, syllables or words matching auditory stimuli played through headphones (see Figure 1a). Thus, the game provides, and documents, unique levels according to in-the-moment skill. The adaptation algorithm of the game ensures that the student receives 20% of trials as challenge and 80% as mastery, based on the individual player's previous performance. Thus, a child has the opportunity to progress to more difficult items, if and when, they demonstrate mastery of more foundational content. In addition to curricular play, i.e., matching audio to letters, syllables or words, Graphogame also

(a)

(b)

FIGURE 1
Screenshots of GraphoGame. **(a)** Curricular play: a trial. **(b)** Extra-curricular: the inventory.

contains extra-curricular elements such as a stickerbook and a shop inventory (see Figure 1b), where the player may spend coins earned in curricular play. GraphoGame has an evidence-base exploring its efficacy (McTigue et al., 2019).

A previous study (name deleted to maintain the integrity of the review process) scrutinized data from 137 6-year old children playing Graphogame over a 25 week period. Progress data was extracted at 5-week intervals throughout the 25 weeks, being operationalized as the number of word items known by the last play session within each time period. This yielded individual growth curves for each child over the 25 week period, with five measurement occasions, which was also compared to a traditional one-time multi-component screen administered at school-entry, which measured letter-sound knowledge, rapid automatized naming two tests of phonemic awareness and also took into consideration family risk. Latent growth curve analyses showed that variation in progress trajectories explained variation in literacy performance at the end of the academic year, post gameplay, to a greater extent than risk status at school entry, as measured by the one-time screening tool. This work pointed to the potential of using dynamic gameplay data to improve prediction accuracy, however it still only used a small fraction of the player data collected.

1.1.3 Early identification utilizing machine learning

Analysis of the "big data" that log files generate necessitates a reconsideration of the type of analysis methods most appropriate for longitudinal prediction. Machine learning (ML) is a more powerful approach that has more capacity than traditional regression analysis to process and make sense of extensive data sets. The latter approach assumes that the risk outcome is determined by linearly combining a set of predictor variables, imposing very strong constraints on how the predictor variables are allowed to interact in the prediction. In contrast, ML encompasses a large variety of flexible ways to identify and non-linearly combine large sets of predictors.

ML is a form of artificial intelligence which aims to improve the performance of a task, in this case, accurate prediction of reading difficulty risk, through a computational training experience (Jordan and Mitchell, 2015). Part of the data collected is designated as the training set for a ML algorithm, using the training experience to tune the algorithm for optimal prediction accuracy. The performance of the trained algorithm can then be validated out-of-sample, against the remainder of the dataset. The use of ML in reading research is still in its infancy, although a recent special issue in this journal (Erbeli and Wagner, 2023) featured studies on the topic of risk prediction that are of relevance to the work reported here. Two studies directly compared the predictive strength of ML approaches with more traditional logistic regression approaches to risk detection, using first/second grade single task emergent literacy performance measures as their input, and third/fourth grade literacy indicators as their outcome measures. Using random forest (Erbeli et al., 2023) and classification and regression tree (Gutirrez et al., 2023) ML models, respectively, the strength of prediction across ML and logistic regression approaches was comparable. In an additional study by Psyridou et al. (2023), a neural network model was compared to linear and mixture models in predicting reading difficulties at Grade 9, from 16 kindergarten predictor variables. In this study, while all three methods worked well, the neural network model appeared to be the most accurate. In the accompanying editorial, Erbeli and Wagner concluded that, "If there is to be a substantial improvement in prediction from the machine learning approach, it is more likely to happen when richer sources of data are incorporated" (Erbeli and Wagner, 2023, p. 4). Included in this, they suggest moving away from single tasks, or data measured over a relatively short time window, rather applying ML to activities such as daily classroom learning within intervention activities.

1.2 Present study

The study presented here goes beyond the status quo in key ways. Firstly, we broaden the scope of gameplay data used

within predictive models, beyond the grainsize of what has been previously reported in reading prediction research. Through more comprehensive data mining and collaboration with the game developers we were able to extract log data from gameplay, including time spent and activities undertaken in extra curricular gameplay, for example, visiting the shop to spend rewards. The rationale for the latter was the accrued evidence of a relationship between task-avoidant behavior and reading skill (Syal and Torppa, 2019). While extra-curricular activity in a digital game cannot be labeled as task-avoidant behavior *per se*, individual differences in the amount of time spent in extra-curricular activity relative to time spent in literacy-related gameplay could be an indicator of relative engagement, and a more direct measure as compared to adult reports via questionnaire, which is a commonly used approach. In addition, we build on the nascent body of work utilizing machine learning for reading difficulty risk detection, providing an innovative application via discrete, instructional gameplay data. We thus explore extensive process data yielded from a sample of 1,676 6-year-old students playing the digital reading game Graphogame for 5 weeks. *In this we ask; Based on 10-min daily gameplay for 5 weeks, how accurately can machine learning methods identify which school starters are at risk of future reading difficulties?*

2 Material and methods

2.1 Data

We next describe the data processing flow, as depicted with blocks (a)–(f) in Figure 2. All data processing, model training and evaluation analyses were performed in the R (R Core Team, 2022) environment.

2.1.1 Participants

At the beginning of the school year in 2021 all first-grade teachers in a southern, urban municipality of Norway, were invited to participate in the study. The study participation involved letting students play the literacy game GraphoGame for ten minutes daily in class, using iPads, over the course of 5 weeks. Working with GraphoGame was already part of the curriculum in the majority of schools. The teachers who participated belonged to a wide variety of schools in the municipality, and we deem the sample of participants to be representative of the population. Among the 6,209 first grade students in the municipality, 1,676 (27%) participated in the present study [block (a) in Figure 2]. Students with fewer than 5 days of gameplay were excluded from the dataset, with 1,640 remaining students (52.7% boys). Mean age at the start of gameplay (September/October 2021) was 6.25 years.

2.1.2 Gameplay data

As part of this study, extended logging functionality was added to Graphogame. All game events were timestamped and saved to a Google Cloud Platform server. For curricular gameplay the stimulus and the distractors were registered, together with screen coordinates, the chosen trial reply and its reaction time. All extra-curricular activity was also logged, such as the spending of gold and silver coins (earned in curricular gameplay) in the shop, visits to the hairdresser, and stickerbook activity. This resulted in a total of over 1 million log files, formatted according to the JSON standard (Pezoa et al., 2016).

The richness and complexity of gameplay data was handled by relying on documented domains of prediction for reading difficulties to construct from the process data a set of 96 features to identify future struggling readers [see block (b) in Figure 2]. The features can be linked and grouped related to dimensions empirically shown to be of importance for predicting reading difficulties, dimensions also underpinning the one-time multi-component assessments pointed to above. These dimensions include (a) students' effort shown in time on task and completion (Robson et al., 2020), (b) the accuracy of the performance (Perfetti and Hogaboam, 1975), (c) aspects of recall and timing related to reading development (Swanson and Howell, 2001; Stanovich, 1981), and finally (d) aspects of extra curricular activity, that could represent task avoidance (Syal and Torppa, 2019) (see Figure 3). Among the 96 features, four were considered redundant, due to being highly correlated with some of the remaining features. Hence, the final feature set contained 92 features [block (c) in Figure 2]. Supplementary Table S1 contains explanations for the full list of features.

2.1.3 National screening test

Our main goal in the present study was to evaluate the potential of machine learning analyses of gameplay features in predicting whether a student will be classified as a weak reader at the end of first grade. Weak reading skill is operationalized as falling below the threshold on a pen- and paper based, group-administered national screening assessment conducted at the end of the first school year (in April/May 2022). The screening test has been psychometrically validated (Norwegian Directorate for Education and Training, 2018, 2015; Walgermo et al., 2021) in large national samples and has been used as outcome measure in several large-scale randomized intervention studies (Lundetræ et al., 2017; Solheim et al., 2018).

Although mandatory, the screening test was not completed by about 5% of the students for which gameplay features was constructed. These students were either exempted from taking the test, or had moved to a different municipality.

Hence our final dataset [see Block (f) in Figure 2] was comprised of $n = 1562$ students that all played a minimal number of trials and completed the reading screening test. In addition, we added age and gender [see Block (d) in Figure 2] to this dataset. The screening is intended to identify 15–20% of the weakest readers, so that these students may benefit from extra resources during the first years of schooling. The screening test is intended to provide at-risk classification on the following five aspects of literacy: letter knowledge, spelling, phonological analysis, word reading, and sentence reading comprehension. The proportion of at risk students in each aspect was 17% for letter knowledge, 22% for spelling, 14% for phonological analysis, 21% for word reading, and 19% for sentence reading comprehension. Being at risk with respect to a particular aspect is positively associated with the likelihood of being at risk with respect to the other aspects. Among

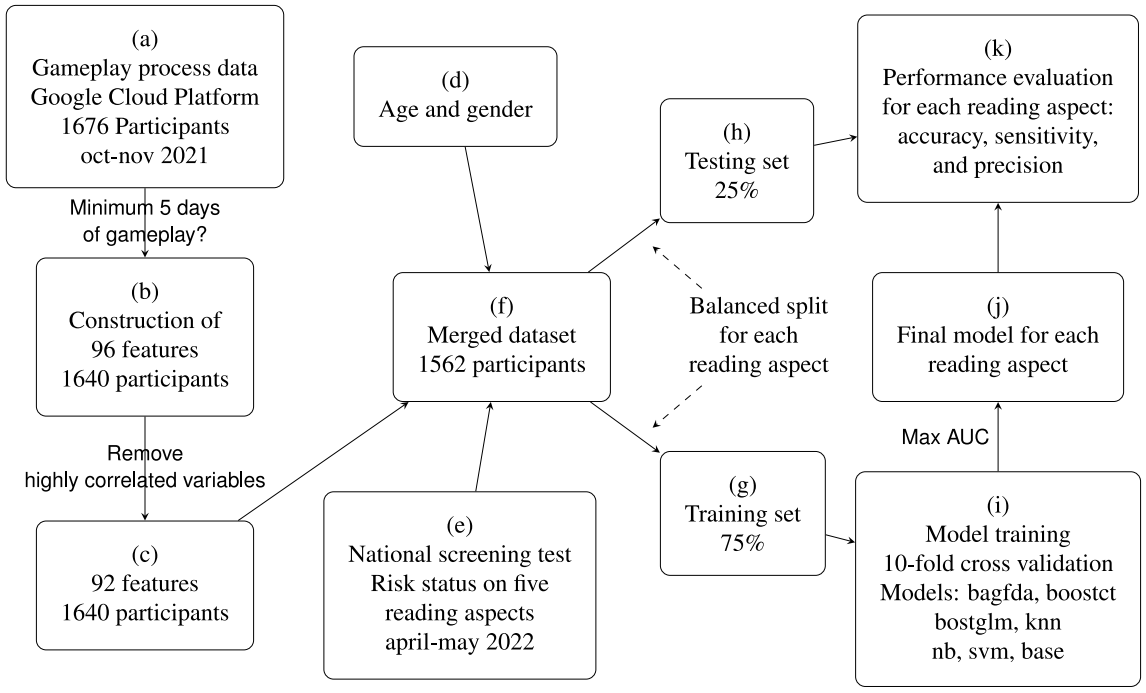

FIGURE 2
Data processing and model training workflow.

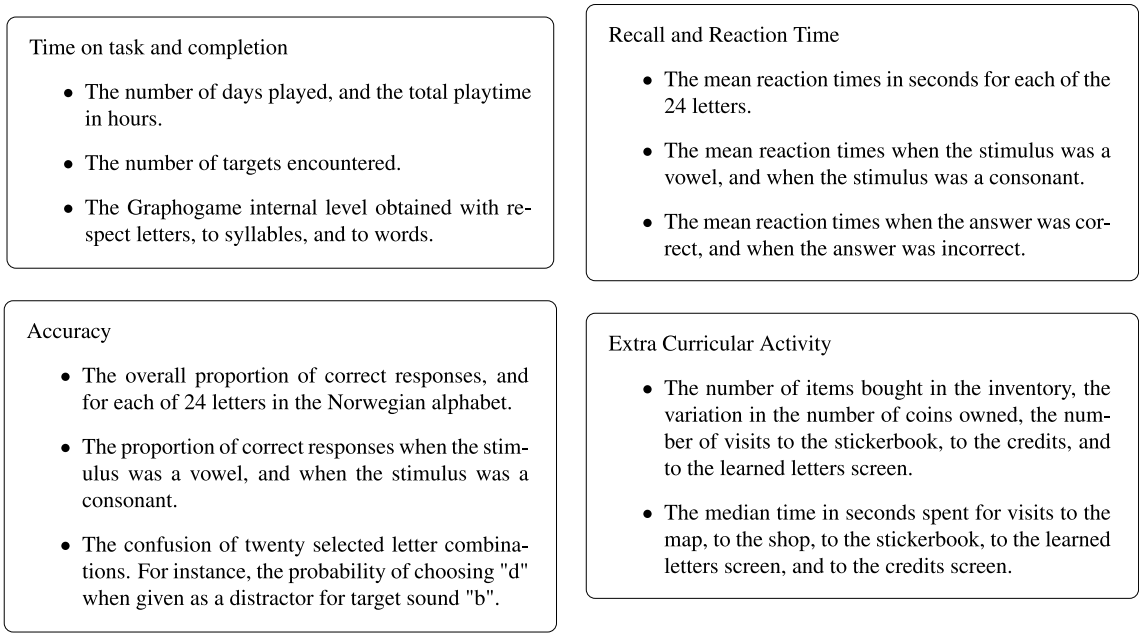

FIGURE 3
Feature groups considered in the present study arranged according to dimensions predicting reading difficulties.

the participants, the Pearson correlation values between test scores varied from moderate (0.54 between letter knowledge and sentence reading) to strong (0.78 between letter knowledge and phonological awareness).

2.2 Model training

We next describe model training as visualized by blocks (g), (i), and (j) in Figure 2. Our task is to classify school starters

according to whether they will fall below the threshold on the end-of-year screening test. As depicted in Figure 2 the merged data was randomly split into testing and training sets containing 25 and 75% of the observations, respectively. The split was balanced, so that the same percentage of at-risk students was retained in the test and training sets. This split was done separately for each of the five at-risk reading aspects. We emphasize that the testing set was a hold-out sample, used exclusively for final evaluation of the chosen binary classifier [Block (k) in Figure 2].

Block (i) in Figure 2 represents the next and crucial step of training models on the training set. Given the lack of prior studies using machine learning to predict reading test scores from gameplay data, no clear candidate for classification algorithm was available a priori. We therefore selected a set of machine learning classifiers to span a range of machine learning approaches in order to identify an appropriate method. The methods include probabilistic and non-probabilistic approaches, and both classical (knn, nb, svm) and recent (bagging and boosting) methods. In addition, as a benchmark we included a simple baseline logistic model. Tuning parameters for each method were determined by the default choice implemented in the caret package (Kuhn, 2022). We here shortly describe the methods in the present study.

bagfda Bagged flexible discriminant analysis is a bagged ensemble algorithm. Flexible discriminant analysis is a non-linear, non-parametric generalization of linear discriminant analysis using multiple adaptive regression splines (Friedman, 1991). The tuning parameter was the number of prunes: 2, 9, 17.

boostct Boosted classification trees (Friedman, 2001) is an ensemble method, where weak classification trees are combined. The tuning grid was constructed from max depth 1, 2, 3 and iterations 50, 100, 150.

boostglm Boosted Generalized Linear Model is also an ensemble method, where generalized linear models are combined. No pruning was used, and the mstop parameter was varied across 50, 100, 150.

knn k-Nearest Neighbors (Altman, 1992) is a non-linear non-parametric method were a player is being assigned to risk status if a plurality of its k nearest neighbors have risk status. The tuning parameters were $k = 5, 7, 9$.

nb Naive Bayes (Murphy, 2012) is a non-linear probabilistic method based on Bayes' Theorem. A tuning parameter determined whether kernel density estimation was employed or not.

svm Support vector machine (Cortes and Vapnik, 1995) is a linear non-probabilistic classifier, which is built by caret with a default cost parameter $c = 1$.

base As baseline model, we fitted a logistic regression model (Cox, 1958) with a simple set of predictors we deemed important: proportion of correct responses, mean reaction time, level attained on syllables, and number of days.

A cross- validation scheme with 10-folds (10-CV), repeated five times, was used to evaluate the performance of the methods. During resampling, to mitigate the issue of class imbalance, we used up-sampling so that in each fold the at-risk class was oversampled.

For each model, parameter tuning was done by optimizing the Area Under the operating characteristic Curve (AUC) (Hanley and McNeil, 1982). The AUC value ranges from 0.5 (random classifier) to 1.0 (perfect classifier). Among the seven candidate models, a final model [block (j) in Figure 2] was determined by maximizing the AUC value.

2.3 Evaluation

The chosen model was evaluated by predicting at-risk status for each participant in the testing set, and comparing it to the participant's true at-risk status. The prediction results may be tabulated in a confusion matrix

		Screening test	
		No risk	Risk
Prediction	No risk	TN	FN
	Risk	FP	TP

containing the frequencies of true negatives (TN), false negatives (FN), false positives (FP), and true positives (TP).

There are many metrics that are used to evaluate trained classifiers. In the present study our main objective is to identify as many at-risk students as possible, while trying to keep the number of false positives as low as possible. Early detection of students that are at risk of obtaining low scores on the national screening tests is more important than misjudging a student that will not score low on the screening test. In our evaluation we therefore first and foremost focus on the *sensitivity* of the classifier, i.e., the proportion of at-risk students that are correctly classified, is of vital importance.

$$Sensitivity = \frac{TP}{TP + FN}.$$

All else equal, it is desirable to achieve as high sensitivity as possible.

The sensitivity must however be balanced against other aspects of classification performance. The *precision* gives the proportion of correctly classified at-risk students, relative to the total number of students classified as at-risk:

$$Precision = \frac{TP}{TP + FP}.$$

All else equal, it is desirable to achieve as high precision as possible.

Finally, we also calculate the overall *accuracy* of the classifier,

$$Accuracy = \frac{TP + TN}{TN + FN + FP + TP}.$$

All else equal, it is desirable to achieve as high accuracy as possible.

3 Results

3.1 Descriptive statistics for feature set

In Supplementary Table S2, are given the univariate descriptive statistiscs for each of the unstandardized features.

3.2 Final model selection

For each of the five reading measures, the final model was chosen based on the mean AUC value obtained from cross-validation. In Figure 4 these values are plotted, and it is seen that models perform quite similarly, with the exception of knn, which performed worse with relatively low AUC values. In general, bagfda and boostglm were found to perform best, with bagfda chosen as the final model for aspects letter knowledge, word reading and sentence reading, while boostglm provided the final model for aspects phonological analysis and spelling.

3.3 Precision, sensitivity, and accuracy

The confusion matrices obtained when applying the chosen classifier (boostglm or bagfda) and the baseline model for the testing set are given in Table 1. Based on these matrices, we calculate for each of the five aspects of reading sensitivity, precision and accuracy, with the results given in Table 2.

Sensitivity is of primal concern when detecting at risk students, and it is seen that the baseline model has poor performance with respect to this measure. Baseline sensitivity is highest for the spelling outcome. Still, only 37% of students at risk for poor spelling performance were detected by the baseline. For the other four aspects of reading, baseline model at-risk detection rates were 26% or lower. In contrast, bagfda and boostglm models achieved far higher sensitivity values, ranging from 72% (letter knowledge) to 80% (word reading). With respect to the secondary performance measure, precision, the baseline model uniformly outperformed the machine learning models. In all five aspects, more than half of the baseline model flagged participants were in fact at risk on the screening test. However, the baseline model generally flagged too few students of being at risk. Concerning the measure of accuracy, flagging few students will inevitably result in accuracy coming close to true proportion of at-risk students. For instance, with respect to letter knowledge, 83% of students were not at risk, so that lazily classifying all students as not being at risk will reach 83% accuracy.

A general pattern observed in Table 1 is that the far better sensitivity values obtained by the advanced models relative to the baseline model come at the cost of increasing the number of false positives. That is, although the advanced models detect far more of the at-risk students than does the baseline model, these models also falsely flag many students who are not at risk. For instance, in sentence reading 72 of the 386 students were at risk. The baseline model flagged only 31 students as being at risk, with 18 of these being truly at risk. The bagfda model flagged 119 students, with 55 being truly at risk. So we may say that the cost of reaching a sensitivity of 76.4% (55/72) is a low precision of 46.2% (55/119).

4 Discussion and recommendations

4.1 Discussion of findings

Using both innovations in stealth assessment and machine learning, this study set out to determine how accurately data from 5 weeks of digital game play could identify the school starters at risk of becoming struggling readers. A priority in our model testing was optimization of test sensitivity—i.e., correctly classifying as many of the at-risk students as possible. Our findings strongly affirmed the value of this approach. Using a set of 92 features of gameplay, representing the classes of time on task, accuracy of performance, recall and reaction time, as well as extra-curricular activity, and testing a series of machine learning models, the bagfda model achieved detection sensitivity of 0.80 for word reading. Across five literacy outcome measures—letter knowledge, phonological analysis, spelling, word reading and sentence reading—the sensitivity to correctly classify at-risk was between 0.72 and 0.80; the overall accuracy of the classification ranged from 0.72 to 0.82.

This level of prediction sensitivity and accuracy compares favorably to existing studies using single time-point assessments with a similar sample and prediction time frame, for example the Norwegian student sample reported in Solheim et al. (2021). Across 918 first graders Solheim et al. (2021) reported end of grade 1 reading fluency risk prediction rates equating to a sensitivity of 0.279, and accuracy of 0.62. However, it is important to note that Solheim et al. (2021) reported a higher level of precision, 0.89, compared to any of the models here (range 0.29–0.57). In a separate sample using a similar screening assessment, though with additional measures including rhythm awareness (Lundetræ and Thomson, 2017), the screening tool administered at school entry and used to predict reading and spelling at the end of grade 1 also had lower sensitivity (0.35 for word reading and 0.49 for spelling), but higher overall accuracy (0.89 for word reading and 0.92 for spelling). It should remarked, however, that in both the study of Solheim et al. (2021), and Lundetræ and Thomson (2017), data were collected in a highly controlled way, i.e., administered one-by-one at school start and researcher led group-administered assessment in the end of year 1, as part of a large RCT. In the present study both the initiation of gameplay and assessment was maintained by the teacher alone, i.e., with a consequently larger risk of reliability issues. Also, the performance metrics in these studies were calculated in-sample, i.e., from the same sample that were used to train the binary classifier. Hence, the classifier may be overfitted to the data at hand, and prediction performance measures may be biased.

From Table 2, we see that the optimal ML model, which had the highest precision, as well as good accuracy and sensitivity was for the spelling aspect outcome, relative to the other aspects. A plausible reason for this could be that this specific aspect, which involves the spelling of single words with regular orthographic patterns, involves more fine-grained processing of a lot more letters than is the case for the items of the other aspects. Put another way, to get a correct score of one spelling item, the student has to actively and correctly generate (as opposed to the *recognition* required in reading) multiple letter-sound combinations, to a degree not demanded from the other aspects. Add to this, that mastery of the game content involved when playing over the first 5 weeks of school, i.e., mainly associating sounds with letters, is a prerequisite for the mastery of spelling. Therefore, much of the data for the ML models is highly relevant for this aspect.

This study adds to a growing body of large-scale studies that demonstrate the value and further potential of machine learning models as an important tool in reading difficulty prediction

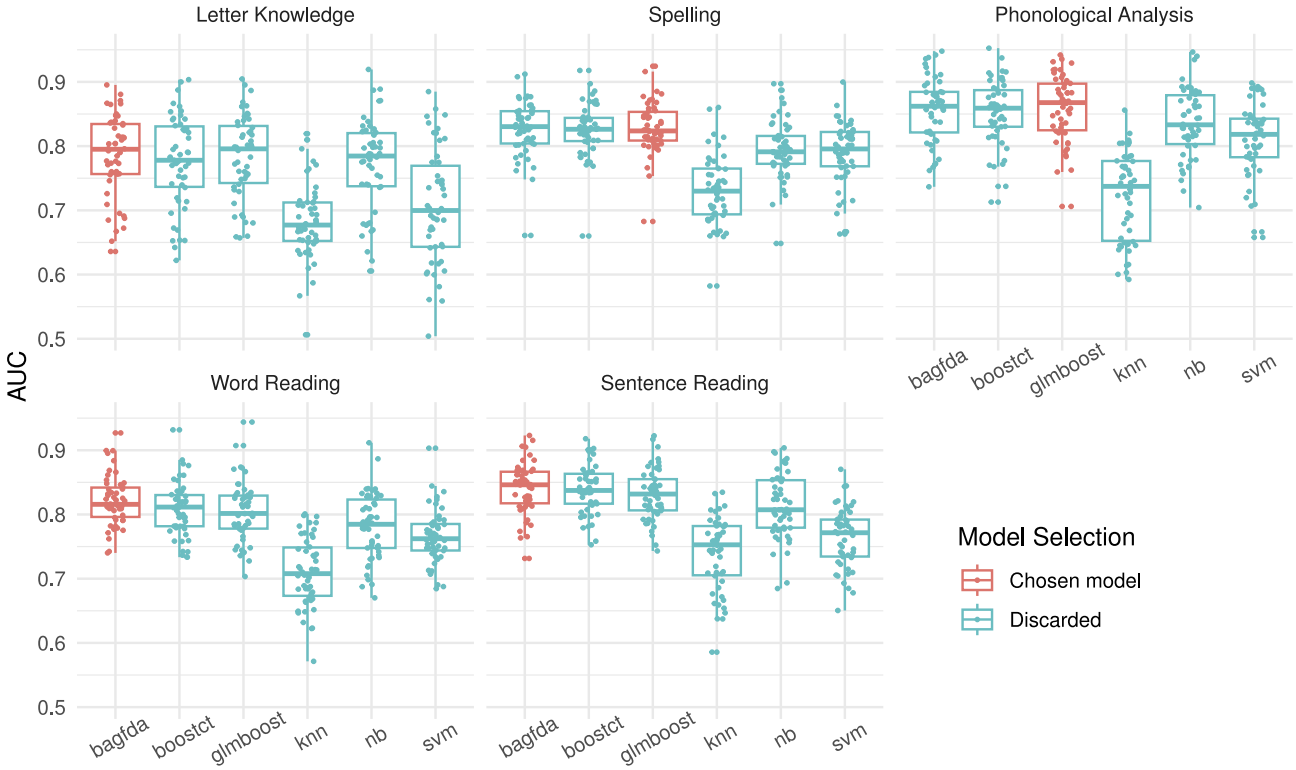

FIGURE 4
AUC values obtained under cross-validation for each of five aspects of reading.

(Erbeli and Wagner, 2023). Previous studies comparing the predictive accuracy of two distinct machine learning models compared to logistic regression approaches for detecting difficulties in the third (Erbeli et al., 2023) or fourth grade (Gutirrez et al., 2023) from first and second grade measures reported comparable accuracy across methods. However, Psyridou et al. (2023), using neural network models to predict Grade 9 reading fluency and comprehension difficulty from seventeen kindergarten age variables found that the neural network model was more accurate than either linear or mixture models. Focusing on shorter-range prediction across the first school year, we report similar levels of prediction accuracy in the current study, using an alternative ML approach. Given the lack of an established best ML approach, here we decided to train and evaluate more than one machine learning model. In-sample cross-validation performance was quite similar (Figure 4) for three of the most advanced models (bagfda, boostct, and boostglm) that were based on bagging or boosting, and better in general than the performance the simpler models knn, nb, and svm. Bagging and boosting are ensemble learning algorithms where the final classification is derived from combining results from several simple models (Kuncheva, 2014). Findings from the present study confirm that ensemble algorithms perform well when trained on educational process data. In addition, the use of readily-collected gameplay data for prediction, in contrast to the time-intensive administration of multiple behavioral assessment measures offers a more scalable model for widespread screening.

4.2 Limitations

While the findings here demonstrate the exciting promise of using complex datasets to better predict the trajectory of complex behaviors such as reading development, prediction also requires sensitive outcome measures of reading. While the advantage of using the Norwegian national screening test as an outcome measure was its robust development process, ease of mass administration and assessment of five complementary aspects of reading and spelling (Norwegian Directorate for Education and Training, 2018, 2015), a screening measure is by nature not comprehensive, and also not designed to fully separate out performance at the higher end of ability; future work should validate the current approach with more fine-grained outcomes of reading performance. Equally, Graphogame has been designed as a reading intervention, as opposed to being specifically designed as a stealth assessment. In some respects, this makes the game an ideal form of assessment, as it allows ecological capture of the very learning process we are trying to predict. However, designing the game with assessment also in mind may allow increased sensitivity, through more planful exposure to gameplay items known from previous research to differentiate struggling from non-struggling readers, for example reading and spelling of consonant clusters, or to identify the optimal adaptation algorithm for struggling readers.

In the present study we extracted features from the process data in a rather manual way, by using expert knowledge

TABLE 1 Confusion matrices for chosen and baseline models when applied to testing set for five aspects of reading.

Outcome							
Letter knowledge			bagfda			Baseline	
			Screening test			Screening test	
			No risk	Risk		No risk	Risk
	Prediction	No risk	247	18	Prediction	313	50
		Risk	73	47		7	15
Spelling			boostglm			Baseline	
			Screening test			Screening test	
			No risk	Risk		No risk	Risk
	Prediction	No risk	254	21	Prediction	297	56
		Risk	49	65		6	30
Phonological analysis			boostglm			Baseline	
			Screening test			Screening test	
			No risk	Risk		No risk	Risk
	Prediction	No risk	236	12	Prediction	326	48
		Risk	97	41		7	5
Word reading			bagfda			Baseline	
			Screening test			Screening test	
			No risk	Risk		No risk	Risk
	Prediction	No risk	230	16	Prediction	288	59
		Risk	74	63		16	20
Sentence reading			bagfda			Baseline	
			Screening test			Screening test	
			No risk	Risk		No risk	Risk
	Prediction	No risk	250	17	Prediction	304	55
		Risk	64	55		10	17

TABLE 2 Sensitivity, precision and accuracy for chosen and baseline model for each of five reading aspects.

Outcome	Model	Sensitivity	Precision	Accuracy
Letter knowledge	bagfda	0.723	0.392	0.764
	Baseline	0.231	0.682	0.852
Spelling	boostglm	0.756	0.570	0.820
	Baseline	0.349	0.833	0.841
Phonological analysis	boostglm	0.774	0.297	0.718
	Baseline	0.094	0.417	0.858
Word reading	bagfda	0.797	0.460	0.765
	Baseline	0.253	0.556	0.804
Sentence reading	bagfda	0.764	0.462	0.790
	Baseline	0.236	0.630	0.832

to connect gameplay behavior to feature extraction. While this captures large aspects of gameplay, it does so in a rather limited form. Dynamic day-to-day behavior that may be indicative of future learning difficulties may not be registered sufficiently well by the 92 features investigated in this study.

Another limitation pertains to the inherent uncertainty in sensitivity and other classification measures that stem from a finite sample size. Although the number of participants in the present study is comparable or in many cases larger than in comparable studies, the splitting of data into training and testing sets means that the testing set contains no more than 380–390 participants. Therefore, there is some statistical uncertainty in, e.g., the sensitivity values reported here.

4.3 Recommendations for future work

As the present study has shown promising results regarding identifying the optimal ML model to apply for this kind of data, a pressing next step is to investigate how to provide the optimal input for the machine learning algorithms. In this, two strands of investigation should be pursued.

Firstly, it is a pertinent question whether the features that serve as input may be optimized. In the present study we have created features that can be linked to hypothetical/empirical factors of reading difficulties. It remains, however, to investigate what gameplay information carries the most information. We do not know the relative importance of the features of these classes, with the predictive utility of specific error patterns, and extra curricular activity particularly unexplored to date. It is an empirical question whether these one or several of these classes of features can be extended—or detailed—to improve the prediction.

Secondly, future studies should consider applying deep learning approaches (Aggarwal, 2018) combined with refined construction of features. This involves building abstract gameplay representation vectors to be used as input for neural network training. This will come at the cost of abandoning human understandable features as used in the present study, but with the potential benefit of more accurate prediction.

Data availability statement

R code for the analyses presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/he2tc/?view_only=dafcfe43a19d4d00bcafd0dcca3a2549. The dataset is available upon request.

Ethics statement

The studies involving humans were approved by Norwegian Data Protection Authority. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

NF: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. PU: Conceptualization, Funding acquisition, Project administration, Writing – original draft, Writing – review & editing. SG: Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. JT: Investigation, Methodology, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Research Council of Norway, project GAMEPLAY (321047).

Acknowledgments

We would like to thank the students, teachers, and administrators of the Oslo municipality, who contributed to this study. Thanks to Division Director for Service Development, Digitalization, and Analytics, Trond Ingebretsen, and special advisor Kjersti Bjønnes, both Oslo municipality, for project administration, facilitating data collection and implementing the solution.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2024.1487694/full#supplementary-material

References

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* 46, 175–185. doi: 10.1080/00031305.1992.10475879

Caravolas, M., Lervåg, A., Defior, S., Seidlová Málková, G., and Hulme, C. (2013). Different patterns, but equivalent predictors, of growth in reading in consistent and inconsistent orthographies. *Psychol. Sci.* 24, 1398–1407. doi: 10.1177/0956797612473122

Cortes, C., and Vapnik, V. (1995). Support vector machine. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cox, D. R. (1958). The regression analysis of binary sequences. *J. R. Stat. Soc.* 20, 215–232. doi: 10.1111/j.2517-6161.1958.tb00292.x

DeWalt, D. A., Berkman, N. D., Sheridan, S., Lohr, K. N., and Pignone, M. P. (2004). Literacy and health outcomes. *J. Gen. Intern. Med.* 19, 1228–1239. doi: 10.1111/j.1525-1497.2004.40153.x

Dixon, C., Oxley, E., Nash, H., and Gellert, A. S. (2022). Does dynamic assessment offer an alternative approach to identifying reading disorder? a systematic review. *J. Learn. Disabil.* 56, 423–439. doi: 10.1177/00222194221117510

Dumas, D., McNeish, D., and Greene, J. A. (2020). Dynamic measurement: a theoretical–psychometric paradigm for modern educational psychology. *Educ. Psychol.* 55, 88–105. doi: 10.1080/00461520.2020.1744150

Erbeli, F., He, K., Cheek, C., Rice, M., and Qian, X. (2023). Exploring the machine learning paradigm in determining risk for reading disability. *Sci. Stud. Read.* 27, 5–20. doi: 10.1080/10888438.2022.2115914

Erbeli, F., and Wagner, R. K. (2023). Advancements in identification and risk prediction of reading disabilities. *Sci. Stud. Read.* 27, 1–4. doi: 10.1080/10888438.2022.2146508

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Stat.* 19, 1–67. doi: 10.1214/aos/1176347963

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Grigorenko, E. L., and Sternberg, R. J. (1998). Dynamic testing. *Psychol. Bull.* 124, 75–111. doi: 10.1037/0033-2909.124.1.75

Gutièrrez, N., Rigobon, V. M. R., Marencin, N. C., Edwards, A. A., Steacy, L. M., and Compton, D. L. (2023). Early prediction of reading risk in fourth grade: a combined latent class analysis and classification tree approach. *Sci. Stud. Read.* 27, 21–38. doi: 10.1080/10888438.2022.2121655

Hanley, J. A., and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143, 29–36. doi: 10.1148/radiology.143.1.7063747

Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260. doi: 10.1126/science.aaa8415

Kuhn, M. (2022). *caret: Classification and Regression Training. R Package Version 6.0–92*.

Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: John Wiley & Sons.

Livingstone, S. (2021). Erasmus medal lecture 2018 ae gm barcelona: realizing children's rights in relation to the digital environment. *Eur. Rev.* 29, 20–33. doi: 10.1017/S106279872000054X

Lovett, M. W., Frijters, J. C., Wolf, M., Steinbach, K. A., Sevcik, R. A., and Morris, R. D. (2017). Early intervention for children at risk for reading disabilities: the impact of grade at intervention and individual differences on intervention outcomes. *J. Educ. Psychol.* 109:889. doi: 10.1037/edu0000181

Lundetræ, K., Solheim, O. J., Schwippert, K., and Uppstad, P. H. (2017). Protocol: 'on track', a group-randomized controlled trial of an early reading intervention. *Int. J. Educ. Res.* 86, 87–95. doi: 10.1016/j.ijer.2017.08.011

Lundetræ, K., and Thomson, J. M. (2017). Rhythm production at school entry as a predictor of poor reading and spelling at the end of first grade. *Read. Writ.* 31, 215–237. doi: 10.1007/s11145-017-9782-9

Lyytinen, H., Erskine, J., Kujala, J., Ojanen, E., and Richardson, U. (2009). In search of a science-based application: a learning tool for reading acquisition. *Scand. J. Psychol.* 50, 668–675. doi: 10.1111/j.1467-9450.2009.00791.x

Lyytinen, H., Ronimus, M., Alanko, A., Poikkeus, A.-M., and Taanila, M. (2007). Early identification of dyslexia and the use of computer game-based practice to support reading acquisition. *Nord. Psychol.* 59, 109–126. doi: 10.1027/1901-2276.59.2.109

McLaughlin, M. J., Speirs, K. E., and Shenassa, E. D. (2014). Reading disability and adult attained education and income: evidence from a 30-year longitudinal study of a population-based sample. *J. Learn. Disabil.* 47, 374–386. doi: 10.1177/0022219412458323

McTigue, E. M., Solheim, O. J., Zimmer, W. K., and Uppstad, P. H. (2019). Critically reviewing graphogame across the world: recommendations and cautions for research and implementation of computer-assisted instruction for word-reading acquisition. *Read. Res. Q.* 55, 45–73. doi: 10.1002/rrq.256

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Norwegian Directorate for Education and Training (2015). *Assessment Test in Reading, Grade 1. Guidelines for Teachers*. Oslo.

Norwegian Directorate for Education and Training (2018). *Framework for Screening Tests at Grades 1–4*. Oslo.

Perfetti, C. A., and Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *J. Educ. Psychol.* 67, 461–469. doi: 10.1037/h0077013

Pezoa, F., Reutter, J. L., Suarez, F., Ugarte, M., and Vrgoč, D. (2016). "Foundations of json schema," in *Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee*, 263–273.

Phillips, B. M., Lonigan, C. J., and Wyatt, M. A. (2009). Predictive validity of the get ready to read! screener: concurrent and long-term relations with reading-related skills. *J. Learn. Disabil.* 42, 133–147. doi: 10.1177/0022194083826209

Psyridou, M., Tolvanen, A., Patel, P., Khanolainen, D., Lerkkanen, M.-K., Poikkeus, A.-M., et al. (2023). Reading difficulties identification: a comparison of neural networks, linear, and mixture models. *Sci. Stud. Read.* 27, 39–66. doi: 10.1080/10888438.2022.2095281

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Robson, D. A., Allen, M. S., and Howard, S. J. (2020). Self-regulation in childhood as a predictor of future outcomes: a meta-analytic review. *Psychol. Bull.* 146:324. doi: 10.1037/bul0000227

Scarborough, H. (2009). "Connecting early language and literacy to later reading (dis) abilities: evidence, theory, and practice," in *Approaching Difficulties in Literacy Development: Assessment, Pedagogy and Programmes*, eds. F. Fletcher-Campbell, J. Soler, and G. Reid (Thousand Oaks, CA: Sage).

Shankweiler, D., and Liberman, I. (eds.). (1989). *Phonology and Reading Disability*. Ann Arbor, MI: University of Michigan Press.

Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., et al. (2021). Maximizing learning without sacrificing the fun: stealth assessment, adaptivity and learning supports in educational games. *J. Comp. Assist. Learn.* 37, 127–141. doi: 10.1111/jcal.12473

Solheim, O. J., Frijters, J. C., Lundetræ, K., and Uppstad, P. H. (2018). Effectiveness of an early reading intervention in a semi-transparent orthography: a group randomised controlled trial. *Learn. Instruct.* 58, 65–79. doi: 10.1016/j.learninstruc.2018.05.004

Solheim, O. J., Torppa, M., Uppstad, P. H., and Lerkkanen, M.-K. (2021). Screening for slow reading acquisition in norway and finland-a quest for context specific predictors. *Scand. J. Educ. Res.* 65, 584–600. doi: 10.1080/00313831.2020.1739130

Stanovich, K. E. (1981). Relationships between word decoding speed, general name-retrieval ability, and reading progress in first-grade children. *J. Educ. Psychol.* 73, 809–815. doi: 10.1037/0022-0663.73.6.809

Stanovich, K. E. (2009). Matthew effects in reading: some consequences of individual differences in the acquisition of literacy. *J. Educ.* 189, 23–55. doi: 10.1177/002205740918900104

Swanson, H. L., and Howell, M. (2001). Working memory, short-term memory, and speech rate as predictors of children's reading performance at different ages. *J. Educ. Psychol.* 93:720. doi: 10.1037/0022-0663.93.4.720

Syal, S., and Torppa, M. (2019). Task-avoidant behaviour and dyslexia: a follow-up from grade 2 to age 20. *Dyslexia* 25, 374–389. doi: 10.1002/dys.1627

Thompson, P. A., Hulme, C., Nash, H. M., Gooch, D., Hayiou-Thomas, E., and Snowling, M. J. (2015). Developmental dyslexia: predicting individual risk. *J. Child Psychol. Psychiatry* 56, 976–987. doi: 10.1111/jcpp.12412

Verhoeven, L., Voeten, M., and Segers, E. (2022). Computer-assisted word reading intervention effects throughout the primary grades: a meta-analysis. *Educ. Res. Rev.* 37:100486. doi: 10.1016/j.edurev.2022.100486

Walgermo, B. R., Uppstad, P. H., Lundetræ, K., Tønnessen, F. E., and Solheim, O. J. (2021). Screening tests of reading: time for a rethink? *Acta Didactica Norden* 15, 1–22. doi: 10.5617/adno.8136

Yang, X., Kuo, L.-J., Ji, X., and McTigue, E. (2018). A critical examination of the relationship among research, theory, and practice: technology and reading instruction. *Comp. Educ.* 125, 62–73. doi: 10.1016/j.compedu.2018.03.009