ARTICLE IN PRESS

Computers & Education xxx (xxxx) xxx

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

# Exploring the Use of Escribo Play Mobile Learning Games to Foster Early Mathematics for Low-Income First-Grade Children

Americo N. Amorim [a,b,*], Lieny Jeon [c,g], Yolanda Abel [c], Stephen Pape [c], Emilia X.S. Albuquerque [b], Monique Soares [b], Vanessa C. Silva [b], Danilo Aguiar [b], José R. Oliveira Neto [b], Claudia Costin [d], Rodrigo L. Rodrigues [e], Mariana Leon [f], Carla A. de Paula [b], Jefferson Lopes [b], Maxsuel S. Silva [b], Maria V. do Nascimento [b], Gabriella A. Patricio [b], Vinícius F. da Silva [b], Raiane Florentino [b]

[a] *New York University, Center for Faculty Advancement, USA*
[b] *Escribo Innovation for Learning, Department of R&D, Brazil*
[c] *Johns Hopkins University School of Education, USA*
[d] *Fundação Getúlio Vargas, Centro de Excelência e Inovação Em Políticas Educacionais, Brazil*
[e] *Federal Rural University of Pernambuco, Department of Education, Brazil*
[f] *Quality Leadership University, Business and Social Sciences School, Panama*
[g] *University of Virginia School of Education and Human Development, USA*

A R T I C L E I N F O

*Keywords:*
Early years education
Games
Mobile learning
Media in education
Teaching/learning strategies

A B S T R A C T

Mobile games can foster the early mathematics skills of children in poverty at home. This randomized controlled trial examines the efficacy of Escribo Play, an evidence-based, game-enhanced early mathematics program with an initial pool of 2980 first-grade students from 267 classrooms in 132 schools located in Brazil. The intervention reached 47% of the children who installed and used the application. Students from 56 classrooms (20%) did not use the application. The intervention retained 51% of those who installed the application. Children of the experimental group who played Escribo Play advanced 2.27 times more than the increase of the control group for number identification, 2.78 times more in count-sequence, 1.73 times more in geometric shape, and 1.41 times more in spatial sense. The overall reduction in disparities between participants from pretest to posttest of the experimental group was 9.1 times higher than the control group, indicating that the intervention reduced learning inequality for those that played the games. The higher gains observed among those children who used Escribo Play may be attributed to its interactive nature, employing evidence-based instructional strategies with engaging and interactive content (i.e., animation and games). The cost per pupil is much lower than reported in 90% of other educational interventions. The application is easy to scale and provides learning gains for first-grade students.

## 1. Introduction

Inequity in learning has been a systematic issue affecting children from underserved families in developed and developing countries

* Corresponding author. New York University, Center for Faculty Advancement, USA.
*E-mail address:* ang6011@nyu.edu (A.N. Amorim).

https://doi.org/10.1016/j.compedu.2023.104759
Received 14 November 2022; Received in revised form 10 February 2023; Accepted 15 February 2023
0360-1315/© 2023 Elsevier Ltd. All rights reserved.

(Schleicher, 2019). The Brazilian education system has been cited as having high levels of learning disparities (Caregnato et al., 2019). According to the Brazilian national assessments, only 47% of the public school fifth-grade students met the mathematics standards (QEdu, 2020). In the ninth grade, the situation is worse. Only 18% met the national standards (QEdu, 2020).

To support children in public schools in Brazil, this study aims to assess the efficacy of an evidence-based early mathematics intervention that provides instructional videos and learning games for first-grade students who live in poverty to study at home. The study was conducted during the COVID-19 pandemic to help reduce the learning issues induced by emergency remote teaching (Hodges et al., 2020). Meta-analyses confirm that structured, in-person interventions and tutoring programs provide substantial early mathematics learning gains for low-income and struggling students (Nelson & McMaster, 2019; Pellegrini et al., 2021). However, it is unclear to what extent online interventions using technology could help children gain in mathematics when used at home.

### 1.1. Previous interventions

To design an intervention that could lead to meaningful learning gains for students, we conducted an extensive review of existing interventions that identify the optimal ingredients, dosage, instructional strategies, and frequency of delivery. Aubrey et al. (2016) examined 29 experimental and quasi-experimental studies of early mathematics programs for PreK and kindergarten that resulted in large effect size ($d = 0.62$) across the studies. The authors found that there was a tendency for programs to produce larger effects when instruction (a) lasted from 120 to 150 min per week, (b) presented content individually to children, and (c) targeted a single content strand.

Nelson and McMaster (2019) examined 34 studies of the effectiveness of early numeracy interventions from Prek to 1st-grade and identified a large effect size ($g = 0.64$). The study also identified ingredients of the interventions. The average duration of the studies reviewed was eight weeks or shorter. Studies also sought to foster knowledge of skills such as cardinality, composition, counting, decomposition, equivalence, magnitude, and set comparison.

Another study of 20 spatial interventions with children aged 0–8 years revealed an effect size of $g = 0.96$ (Yang et al., 2020). This study analyzed effective instructional strategies and found that interventions involving play, games, or hands on activities were more effective than other instructional strategies.

The intervention dosage and frequency of delivery were investigated as well. A meta-analysis of 36 experiments of educational applications (apps) for early learners identified that the average intervention was conducted in 21-min blocks, for a total of 32 sessions, for 87 days (Kim et al., 2021). These apps significantly improved mathematics skills ($g = 0.29$).

Although a 0.29 effect size may be perceived as low when compared to traditional benchmarks (Cohen, 1969), it is more advisable to employ benchmarks specific to the field and based on studies that meet rigorous standards (Lipsey et al., 2012). Examining 314 studies of mathematical trials, Kraft (2020) found that the average effect size was 0.09 for kindergarteners. According to Kraft's schema for interpreting educational interventions, the mean effect size for apps, $g = 0.29$ (Kim et al., 2021), and according to the benchmark criteria, it can be perceived as an easy-to-scale intervention with a large effect and a low cost.

### 1.2. Implementation gaps in the existing interventions

Although the existing body of knowledge indicates that app-based interventions can provide significant early mathematics gains, it is unclear if we can obtain such effects by apps that target the development of multiple early mathematics skills. For example, a recent meta-analysis identified that spatial training interventions, employing concrete materials were more effective than exclusive computer-based instruction (Hawes et al., 2022). Beyond observing the skills, dosage, and delivery mode, it is advisable to consider the challenges that teachers usually face when conducting instruction during the intervention design process.

Interventions that provide intensive teacher training can be effective; however, they are difficult to scale without well-trained coaches/trainers to maintain fidelity (Kraft et al., 2018). Moreover, even if professional development is scaled, interventions often lack cost-effective mechanisms to monitor implementation (Moir, 2018). It is also expensive to provide ongoing coaching for teachers in their implementation. Additionally, it is not common to have a reliable, low-stress, and cost-effective assessment tool to track student progress throughout the year to inform instructional decision-making (Dumas & McNeish, 2017). Fourth, it is not easy to keep children's attention and successful engagement in teacher instruction (Scull & Lo Bianco, 2008).

Early childhood educators have many duties. One of the most important responsibilities is ensuring that students receive high-quality instruction and meet their learning goals. Although changing societal factors is beyond the capacity of early childhood educators, there are classroom factors that have been tied to student achievement. One factor is teachers' instructional strategies encompassing the deliberate instruction of early mathematics concept development (Clements & Sarama, 2008). A second factor is using assessments to monitor student progress and inform policy and practice (Platas et al., 2016). A third is adopting quality learning materials and educational technologies (Dorouka et al., 2020).

Although these challenges were mostly studied in the classroom context, some of them also apply to home-based interventions such as the difficulty to collect reliable data to track implementation and learning. At home, caretakers also lack instructional knowledge, time, and resources to foster learning. Therefore, it is important to find ways to better deliver and monitor instructional interventions that aim to provide supplementary instruction at home.

### 1.3. Potential of technologies for early mathematics instruction

One potential solution to reduce implementation challenges is by using technologies. A meta-analysis of 15 digital-based interventions revealed an effect size of $d = 0.55$ on mathematical learning (Benavides-Varela et al., 2020). Educational apps also have the potential to inform and support teachers' decision-making when they provide learning analytics (Du et al., 2021). Despite such potential, learning analytics may be perceived as superfluous in contexts with multiple standardized assessments that can lead to test anxiety and diminished learning and mental health (von der Embse et al., 2018). Nevertheless, learning analytics may play a relevant role in preschool and kindergarten classes.

There are multiple methods to assess children's learning and development in the literature, including direct assessments and observational assessments. However, each has its limitations when applied to early childhood education (ECE) settings (Benson et al., 2019; Moro & de Souza, 2014; Russo et al., 2019). Without reliable data that can be consolidated, program leaders are often ill-equipped to make relevant curriculum, policy, and instructional decisions. However, game-based stealth assessment has emerged as a promising alternative in ECE settings. It can be a tool to assess learning at scale without the stress associated with standardized tests and the high costs of direct assessments (Shute & Rahimi, 2017).

Stealth assessment data can produce child-level reports for teachers and families and classroom- and school-level reports for teachers and leaders. This feature may be a cost-effective, non-stressful, and reliable way of assessing student development (Shute et al., 2021). These learning analytics support educators and school leaders in further understanding pathways that can help identify why some children might do well at one task but not another.

### *1.4. The present study*

Based on the previous literature findings, an instructional program was created employing digital videos and games to engage children in building their early mathematics content knowledge and skills. The design team compared the mathematics skills and content reported by a previous meta-analysis with the Brazilian curriculum (Ministério da Educação, 2018). It included mathematics skills and content aligned with the Brazilian mandatory learning objectives. The digital videos and games included number identification, count-sequence, set comparison, spatial sense, and geometric shape recognition. The program was designed to be used by first-grade students and offered several features to facilitate its dissemination to low-income families. The application could be run on inexpensive computers, smartphones, or tablets. It employed data compressing and caching methods to reduce bandwidth consumption while downloading instructional games. Parents could download the content with internet access, and the students could use the application when the device was offline.

This intervention offered 20 lessons, each comprised of an instructional video followed by a digital game (see Appendix). All games collected data while the children played for the future development of student's performance reports employing stealth-assessment techniques (Shute et al., 2021). Each lesson could be completed within 30 min to reduce screen time and facilitate usage in low-income households where all family members share the same mobile device. This same software infrastructure was previously employed to develop and deploy an evidence-based early literacy program that demonstrated significant early reading and writing gains in schools serving students from low-income and middle-class backgrounds (Amorim et al., 2022).

Furthermore, there is a lack of standardized assessments for measuring emerging academic skills, such as mathematics, in Brazil. Although some instruments were developed to measure arithmetic knowledge, an extensive literature review could not identify any assessment for foundational skills such as number identification, counting, sequencing, and spatial reasoning. Given the lack of instruments in the Brazilian Portuguese language, one option would be to translate and culturally adapt assessments developed in other countries. However, most assessments for early grade students employ an in-person evaluator, as some students do not have the skills to read, understand, and answer a test alone. During the COVID-19 pandemic school closures, meeting with students in person would have been impossible. The remote application of instruments developed for in-person application was also problematic. Given the implausibility of in-person assessments, the researchers also designed and tested the Gamified Early Mathematics Mobile Assessment (GEMMA) as part of the application implementation in this study.

In developing countries there still exists discussion among practitioners if schools should invest in technology for early learners as "most families are poor and will not be able to use it". Nevertheless, in Brazil 90% of the population have a smartphone. Thus, we sought to examine the extent that the intervention could retain students from low-income families when they are given the opportunity to use tech for learning at home.

### *1.5. Research questions*

This research, then, sought to answer the following questions:

Question one: To what extent can a mobile-based intervention retain student participation of low-income families at home during school closures?

Question two: What are the validity and reliability of the GEMMA assessment which is incorporated within the app?

Question three: To what extent did the mobile-based gamified intervention designed to foster number identification, set comparison, count-sequence, geometric shape recognition, and spatial sense skills affect students' mathematics knowledge and skill compared to the control children that did not receive the program?

## 2. Material and methods

We used a randomized controlled trial (RCT) to assess the learning outcomes of the intervention (Shadish et al., 2002). After the pretest, the students were randomly allocated to the experimental or control group. Randomization was conducted with a spreadsheet containing the list of all classrooms. A random number was generated between 0 and 1 for each classroom using the Microsoft Excel = RAND function. Classrooms with the random number between 0.00 and 0.49 were assigned to the control group while those with the number between 0.50 and 1.00 were assigned to the experimental group.

The experimental group of students received the early mathematics instructional program. To ensure that the control group still had equitable access to learning resources during the unprecedented challenging time of the COVID-19 pandemic, we offered access to the same application for the control group, however, they received games unrelated to the early mathematics skills (i.e., early literacy games).

Following the instructional sessions, all students completed the posttest. After both groups finished the posttest, the children of the experimental group received the early literacy intervention, and the children of the control group received the mathematics sessions. Ultimately, all children received literacy and mathematics interventions to promote learning equity.

### 2.1. *Ethical procedures*

The study procedures were approved by the Secretary of Education of the two participating cities where the public schools were located. The superintendent of a non-profit network of schools also approved the study protocol. The study was conducted as a regular educational practice of the participating schools of the two cities and the non-profit network during the 2020 academic year. Before using the application, the teachers and families consented to data collection, analysis, and dissemination procedures. All participation was voluntary, and participants did not receive any compensation. The families who did not want to participate in the research could still play the games without data collection by selecting the "Play without collecting analytics" feature of the application. Teachers, principals, and superintendents had access to their student's learning analytics. Students that did not play the games did not receive any academic reprimand.

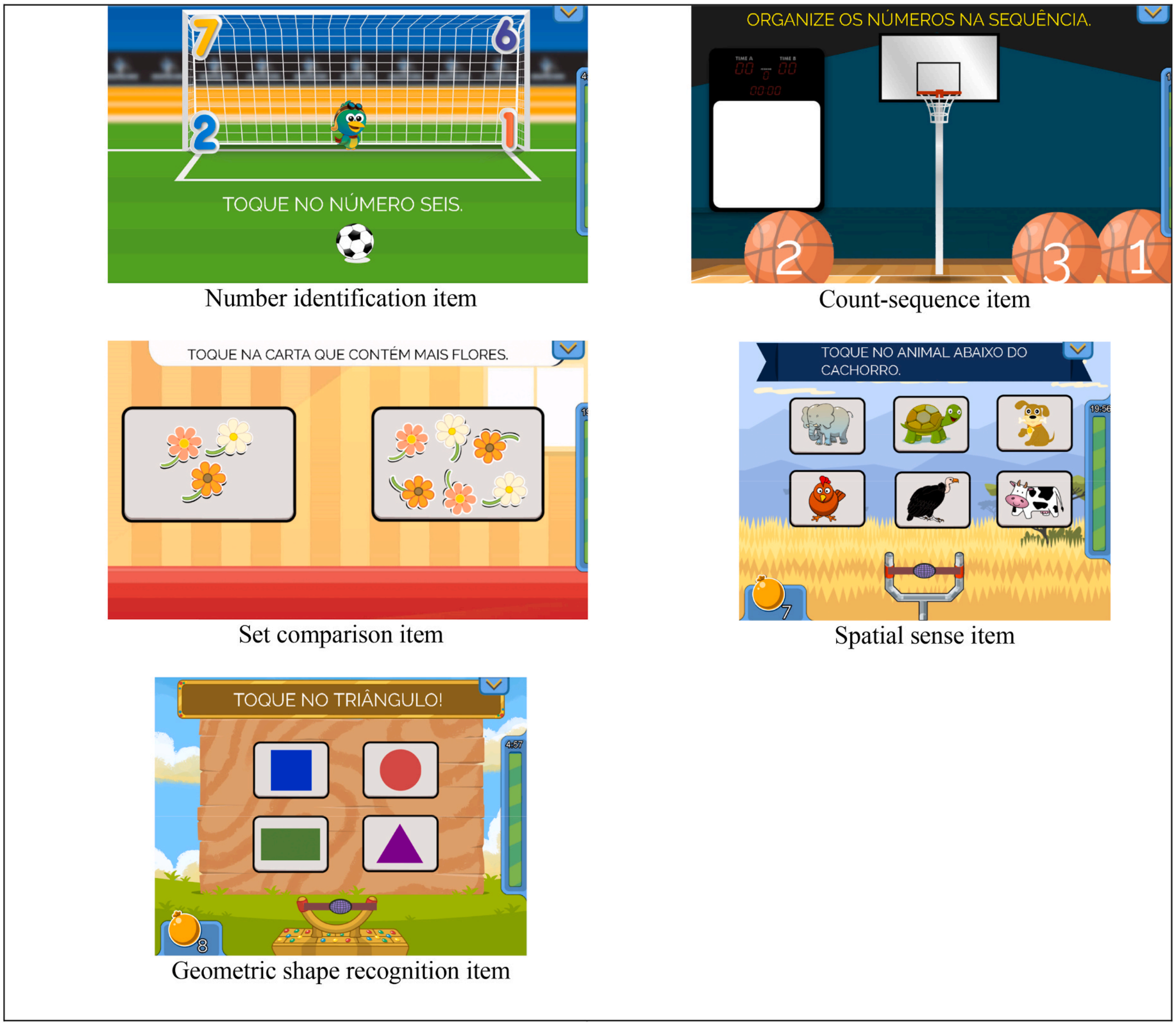


**Fig. 1.** Selected GEMMA assessment items.

### 2.2. Participants

The randomized controlled trial participants were 2980 first-grade students from 267 classrooms in 132 schools. Most of the schools, 123 institutions, were public municipal schools in the southeast region of Brazil, while nine schools were in the northeast region. Most of the schools, 113, were in a city with a medium Human Development Index (HDI), performing 17.85% below its goal defined by the Brazilian quality of education index (QEdu, 2020). Ten schools were in another city with medium HDI, performing 3.27% below their goal (QEdu, 2020). The remaining nine schools were from a non-profit network that served children from families in poverty and were not included in the Brazilian quality goals. All schools remained closed during the entire RCT due to the COVID-19 pandemic.

### 2.3. Instruments

The GEMMA instrument, developed as part of the application implementation, comprises 14 items designed to measure students' knowledge about foundational mathematics skills and concepts, including number identification, count sequence, set comparison, spatial sense, and geometric shapes. The GEMMA assessment looks like a regular digital game to avoid the stress many young children experience when taking standardized assessments. The software plays an audio file in which an actress reads the questions, allowing preschool, kindergarten, and first-grade students who still do not read to 'play' the assessment by themselves. Fig. 1 presents some screenshots of the GEMMA assessment.

As students advance in the assessment, they receive virtual points and stars. When they finish the instrument, they receive a medal depending on how well they played. Students can play the GEMMA assessment multiple times. The GEMMA software stores all students' answers if played without an internet connection and sends them to a secure cloud storage when their device connects to the internet. The GEMMA assessment scores are calculated employing multiple data. First, the assessment provides the number of sessions (NS) that the children tried the assessment until they finish it. This happens as students can open the gamified assessment, answer some items, and close the assessment without concluding it, then open again, starting another session. The more times the assessment is opened until completion indicates that the students may have learned the answers to the items they interacted with multiple times.

Then the software extracts the data collected during the first session that the student effectively concluded the assessment. Such data contains the number of answers (NA) that the student provided in each of the 14 items until he answered correctly. The score for each assessment subtest is then calculated by multiplying the number of sessions by the sum of the number of answers for that subtest items. For example, students may have tried the assessment twice until completion (NS = 2). In the three first items that compose the number identification subtest, they provided three answers for the first item, two for the second, and one for the third (NA = 3 + 2 + 1 = 6).

To deal with the tendency of learning the answers while playing the assessment multiple times, we multiply the number of sessions by the number of answers to obtain the raw subtest score (2 * 6 = 12). Last, we normalized the students' subscores using the min-max procedure (Patro & Sahu, 2015) so that all values stayed between 0 and 1. The higher the number, the better the student's performance. The min-max procedure is repeated for each of the subtests.

The GEMMA instrument was pretested with a sample of 251 children and demonstrated an adequate level of internal consistency ($\alpha = 0.759$). A retest with 95 participants was conducted 21 days after the first application. A paired sample *t*-test indicated that the scores of the students did not differ significantly between these two assessment administrations ($t = 0.145$, $p = .885$, $df = 94$). These preliminary results indicated adequate internal consistency levels and the GEMMA instrument's test-retest reliability.

During the software development, the data collection and storage procedures were assessed using a structured battery of tests developed by the research team. Three researchers played the assessment independently. While playing, they manually registered the items that were scored correctly and incorrectly in a spreadsheet. Each tester played each game four times. It was established that in each of such playing sessions, the tester provided a different answer to each question, thus covering 100% of the possibilities offered by the assessment (which presented four alternatives). After the end of the tests, the spreadsheets filled by the testers were compared to the ones extracted from the application database, which indicated that 100% of the total number of correct and incorrect answers were consistent.

### 2.4. Data collection

The study was conducted during the first year of the COVID-19 pandemic. The student-level intervention was designed and delivered when it became clear that the schools would remain closed until the end of the 2020 academic year. The research team generated login and passwords to use the application for each student from the data provided by the participating schools. The pretest of the early mathematics intervention was conducted in July 2020.

### 2.5. Classrooms were randomized to the experimental and control groups

After the pretest, the experimental group of students received the 20 mathematics instructional sessions to be completed once per day. Each session included one instructional animation that lasted approximately 4 min and explained the target skill of the session. After the animation, the children played the learning game to practice the target skill. The appendix provides information on each session. After delivering the 20 sessions, students were asked to play the posttest GEMMA assessment.

The early mathematical games delivered to the experimental group were designed to foster the same skills assessed by the GEMMA instrument. The instructional sequence was a spiral. The five skills were introduced in a repetitive, spiral-like fashion, building on previous knowledge, and gradually increasing in complexity. This approach fosters the motivation stemming from the games with repetition, reinforcement, and gradual introduction of new concepts and skills, making it easier for students to grasp, retain, and apply the skills. The instructional sequence is detailed in the appendix.

The control classrooms also received 20 games not related to mathematics. During this period, they also received emergency remote teaching of regular mathematics instruction from their classroom teachers. The curriculum used by the schools included the mandatory learning goals established by the Brazilian National Common Core (BNCC). These goals include the skills that were provided to the experimental group: number identification, set comparison, geometric shape recognition, count-sequence, and spatial sense. After the control group students received the literacy component, they played the GEMMA posttest.

### *2.6. Data analysis*

The data were analyzed using the Microsoft Excel, SPSS 26 and R. The total number of participants in each research phase for the experimental and the control groups was computed with a customized spreadsheet to answer question 1 (retention). All lines with missing data were removed (e.g., students who did not complete the pretest, children who did the pretest but did not complete the posttest).

To answer question 2, we examined internal consistency and test-retest reliability. In addition, we conducted the exploratory factor analysis to validate the factor structure of the GEMMA assessment. Two exploratory principal axis factor analyses were conducted on the 14 items with varimax rotation.

To answer question 3, first, we extracted the descriptive data for each subtest score for both groups by employing the SPSS cross-table feature. This table was exported as a spreadsheet. Then we calculated the difference of each group from pretest to posttest using the spreadsheet. Second, we conducted the multilevel analysis in SPSS 26. We input the pretest scores of each subtest and the experimental condition as fixed effects and the classroom where the students were nested as a random parameter. The results obtained in SPSS were compared to those obtained during subsequent analysis conducted in R (version 4.2.2).

We employed the partial eta squared of the group allocation variable as the effect size measure for each subtest. The effect sizes were converted to their equivalent *d* to facilitate discussion and dissemination among educational researchers and practitioners using software provided by Lenhard and Lenhard (2016).

## 3. Results

### *3.1. Question 1 – student retention*

A total of 2980 users were created in the initial database. The logins and passwords were given to the teachers to be delivered to the parents. A total of 1416 students completed the first game, the pretest (911 from the control group and 505 from the experimental group). Although 47% of the children installed and effectively used the app, we do not have data regarding how many of the remaining did not receive the login password from their teachers and thus were prevented from participating.

At the end of the intervention, 726 children fully participated until the end of the program and thus completed the posttest (51% of the 47%). In the control group, 479 children were retained out of the original 911 (52%). The experimental classes retained 247 students out of 505 (49%). These retention rates are similar to other previous home-based interventions (Snell et al., 2022).

**Table 1**
Items of the GEMMA assessment and factor loadings.

| Items/Factors | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Touch on the number 3 | .051 | .013 | **.703** | .009 | .062 |
| Touch on the number 6 | .088 | −.005 | **.870** | .120 | .066 |
| Touch on the number 10 | .009 | .110 | **.822** | −.044 | .090 |
| Throw the balls in the correct sequence (3 1 2) | .179 | **.761** | .057 | .127 | .053 |
| Throw the balls in the correct sequence (5 6 4) | .059 | **.844** | .035 | −.005 | .040 |
| Throw the balls in the correct sequence (8 7 9) | .110 | **.866** | .031 | .107 | .094 |
| Touch on the painting with more flowers (6 or 3) | .106 | .105 | .039 | .061 | **.856** |
| Touch on the painting with less pencils (3 or 5) | .095 | .052 | .170 | −.029 | **.833** |
| Touch on the animal that is BELOW the dog | .098 | .029 | .027 | **.877** | .032 |
| Touch on the animal that is ABOVE object X | .214 | .188 | .045 | **.794** | .000 |
| Aim on the triangle | **.737** | .114 | .136 | .109 | .126 |
| Aim on the circle | **.821** | .154 | .033 | .057 | .113 |
| Aim on the rectangle | **.716** | −.060 | −.006 | .228 | .005 |
| Aim on the square | **.800** | .203 | .026 | .008 | .030 |
| Cronbach's alpha – Pretest | .692 | | | | |
| Cronbach's alpha – Posttest | .694 | | | | |

### 3.2. Question 2 – GEMMA assessment

During the RCT, with complete data from 1416 participants at the pretest and 912 participants at the posttest, we also obtained an adequate level of internal consistency for the GEMMA instrument, $\alpha = 0.70$ at the pretest and .72 at the posttest. Two exploratory principal axis factor analyses were conducted on the 14 items with varimax rotation. The Kaiser–Meyer–Olkin measure indicated an adequate sample for factor analysis with KMO = 0.70 for the pretest and .73 for the posttest. All items had a KMO between 0.58 and 0.80 at the pretest and .60 and .84 at the posttest, all above the suggested criteria of 0.5. An initial extraction indicated five factors with eigenvalues above Kaiser's criterion of 1. Together they explained 63.35% of the variance at the pretest and 69.10% at the posttest. We retained the five factors. Table 1 presents the factors with their rotated loadings for the posttest analysis.

Observing the factor loading of each item, it becomes clear that factor 1 represents the geometric shape recognition skills because the last four items loaded higher on this factor. Factor 2 encompassed the three count-sequence items. Factor 3 included the three number identification items. Factor four loaded higher on the two spatial sense items. Finally, factor five loaded higher on the two set comparison items. The alignment between the assessment items and their underlying constructs with the factors revealed by the student's responses indicates that the GEMMA assessment presented acceptable content and construct validity levels.

### 3.3. Question 3 – intervention efficacy

To examine whether there were differences between the experimental and control groups prior to the intervention, we conducted an independent samples *t*-test. Table 2 shows that the independent samples *t*-test revealed significant differences at pretest between the two groups for all skills, except for geometric shape identification. Because of such differences we included the pretest scores in the multilevel analysis.

Table 3 presents the descriptive statistics of the pretest and posttest scores by mathematics category. The mean differences between the pretest and posttest revealed that the experimental group advanced more than the control group in all subtests with the exception of the set comparison skill. Nevertheless, the standard deviation of the experimental group for set comparison reduced .056 compared to a difference of 0.012 from the control group. These results indicates that experimental group reduced its dispersion 4.67 times more than the control group reduction (0.056/0.012 = 4.67).

The experimental group reduced its dispersion more than the control group for all the other four subtests. Such reduction in dispersion indicates that students from the experimental group were more cohesive in their skills, with fewer students on the extremes (lagging behind our much more advanced). For number identification the differences among the experimental students were reduced 6.88 times more than reduction of the control group (−0.055/0.008). A similar pattern occurred for count-sequence with 3.54 times more and geometric shape with 20.7 times more reduction than the reduction of the control group. For spatial sense, the reduction of the experimental group was 70% more than the reduction of the control group.

Regarding the mean differences, the experimental group displayed a 2.27 times larger increase than the increase of the control group for number identification (0.025/0.011). The experimental students also displayed larger increases than the increase of the control group in count-sequence (2.78 times), geometric shape (1.73 times) and spatial sense (1.41 times) skills. To assess the significance of those differences, multilevel models were employed due to the clustered structure of the data (students within classrooms). Table 4 presents the multilevel models' outcomes. After controlling for the pretest scores, the treatment significantly affected the posttest scores for set comparison, count-sequence, geometric shape, and spatial sense skills.

The multilevel model revealed that the treatment only approached significance for the number identification skill ($p = .088$). Pretest scores had a significant effect on all subtests except set comparison ($p = .522$). Another aspect revealed was that the classroom effects were almost inexistent. Although the students were nested in classrooms, the schools were closed, and students received the intervention at home. The physical distance between teachers and students may have been reflected in a digital distance due to problems stemming from the emergency remote teaching procedures adopted by the schools. Table 5 presents the effect sizes on the five subtests.

## 4. Discussion

### 4.1. Question 1 – student retention

Before the RCT, there was doubt among school leadership regarding the capability of low-income families to use the mobile application to receive instructional sessions. Some educators expressed worries that the families would not be able to download the

**Table 2**
Independent Samples *t*-test at Pretest.

| | *t* | df | Sig. |
|---|---|---|---|
| Number identification | 3.848 | 322.726 | 0.000 |
| Set comparison | −12.328 | 593.672 | 0.000 |
| Count-sequence | 6.838 | 308.805 | 0.000 |
| Geometric shape | −1.623 | 422.655 | 0.105 |
| Spatial sense | 2.911 | 363.433 | 0.004 |

**Table 3**
Descriptive analysis of the early mathematics outcomes.

| | | Control | | | Experimental | | |
|---|---|---|---|---|---|---|---|
| | | Pretest | Posttest | Diff | Pretest | Posttest | Diff |
| Number identification | Mean | .920 | .931 | .011 | .897 | .922 | .025 |
| | Std. Dev. | .047 | .055 | .008 | .087 | .032 | −.055 |
| Set comparison | Mean | .871 | .899 | .029 | .945 | .963 | .018 |
| | Std. Dev. | .088 | .075 | −.012 | .071 | .015 | −.056 |
| Count-sequence | Mean | .953 | .962 | .009 | .906 | .931 | .025 |
| | Std. Dev. | .050 | .063 | .013 | .102 | .057 | −.046 |
| Geometric shape | Mean | .881 | .896 | .015 | .893 | .919 | .026 |
| | Std. Dev. | .087 | .084 | −.003 | .105 | .043 | −.062 |
| Spatial sense | Mean | .940 | .957 | .017 | .918 | .941 | .024 |
| | Std. Dev. | .072 | .035 | −.037 | .107 | .044 | −.063 |

**Table 4**
Multilevel analysis of the Early Mathematics outcomes.

| Fixed Effects | Number identification | | | | Set comparison | | | | Count-sequence | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B | SE | t | Std. B | B | SE | t | Std. B | B | SE | t | Std. B |
| Intercept | .816* | .025 | 31.653 | – | .883* | .024 | 36.146 | | .814* | .029 | 27.598 | – |
| Treatment | −.006+ | .003 | −1.709 | −0.060 | .062* | .005 | 11.863 | 0.430 | −.023* | .004 | −4.800 | −0.180 |
| Pretest | .124* | .027 | 4.460 | 0.170 | .017 | .027 | 0.633 | 0.020 | .154* | .030 | 5.012 | 0.190 |
| **Random Parameters** | | | | | | | | | | | | |
| Residual | .002* | | | | .003* | | | | .003* | | | |
| Classroom | .000 | | | | .000 | | | | .000 | | | |

| Fixed Effects | Geometric shape | | | | Spatial sense | | | |
|---|---|---|---|---|---|---|---|---|
| | B | SE | t | Std. B | B | SE | t | Std. B |
| Intercept | .774* | .025 | 30.838 | | .905* | .015 | 58.496 | |
| Treatment | .021* | .005 | 3.486 | 0.140 | −.014* | .003 | −4.508 | −0.180 |
| Pretest | .138* | .028 | 4.801 | 0.170 | .054** | .016 | 3.333 | 0.120 |
| **Random Parameters** | | | | | | | | |
| Residual | .004* | | | | .001* | | | |
| Classroom | 6.4E-5 | | | | 4.89 | | | |

*Note*. $^*p = .000$, $^{**}p = .001$, $^+p = .088$.

**Table 5**
Effect sizes of the mathematics outcomes.

| | partial $\eta^2$ | $d^+$ |
|---|---|---|
| Number identification | .004 | 0.126 |
| Set comparison | .164 | 0.885 |
| Count-sequence | .031 | 0.357 |
| Geometric shape | .020 | 0.285 |
| Spatial sense | .030 | 0.351 |

*Note:* $^+d$ *obtained by transforming the partial* $\eta^2$.

games to their smartphones due to the lack of widespread internet connectivity. The research team suggested to the school administrations that the potential for benefit to students during the COVID-19 pandemic and the ability for students to use the app offline might have made the potential for its use greater than expectations.

Before investing in new applications and related cloud backend, researchers, developers, and policymakers may consider using an established delivery technology (such as Escribo Play). The relevant requirements for such applications are compatibility with low-cost Android smartphones and low-bandwidth consumption mechanisms. Moreover, offline usage while still collecting learning analytics and playing rich-media content (e.g., video and games) may also be a key factor to consider.

In this study, we found that about 47% of the families installed the application and played the pretest game. This means that about half of families from low-income background had access to the resources necessary to use a mobile-based intervention that can run offline and on low-cost devices and were willing to try this out. Nevertheless, this number can probably be improved as only 211 out of 267 classrooms had at least one student who played. A possible cause for the 56 classes (20%) not having played is that teachers did not distribute the logins and passwords for the families. Increasing the communication with the teachers may be a future improvement to ensure that all students receive access credentials.

The intervention also retained 51% of the students participating until the end. The lack of sufficient mobile devices for all family members may have reduced retention. Improving outreach efforts to demonstrate the potential learning gains for the parents may be essential to increase student retention in future studies and during scale-up efforts.

Another factor is that students may have experienced internet shortages and thus could not download the remaining games, as they were released once per day. Perhaps the application could download all games at once, during the first login. Then the application could release the content for the children according to the delivery sequence and schedule (one game per day). Such change could improve access to the intervention for families that do not have the internet at home.

### 4.2. Question 2 – GEMMA assessment

The internal consistency and test-retest reliability results indicate that the GEMMA assessment was pertinent to a context where there was no adequate assessment of early mathematics skills that could be employed remotely. Our efforts to build this gamified, app-based assessment, and validate in Brazil in Portuguese were fruitful. The factorial analysis showed clear factor structure aligned with the constructs that the instrument was designed to measure.

Nevertheless, replication studies are needed to validate the instrument using various samples. New studies can examine the application of the GEMMA assessment with students playing the assessment in-the school under the supervision of a research assistant and at home to assess eventual discrepancies among those two settings. The instrument also needs to be tested with students from other age groups.

### 4.3. Question 3 – intervention efficacy

Beyond demonstrating the feasibility of delivering game-enhanced educational interventions to children from low-income families at home during school closures, the RCT also identified the learning gains that resulted from such an innovative approach. Although this is the first time an educational intervention was assessed when used by a low-income population in Brazil while students are studying at home during school closures, we can compare its outcomes with previous metanalytical findings of rigorous RCTs conducted in school settings in other countries.

Overall, the children of the experimental group that played Escribo Play experienced strong gains in count-sequence ($d = 0.357$), set-comparison ($d = 0.885$), geometric shape recognition ($d = 0.285$), and spatial sense skills ($d = 0.351$). Those effect sizes are in line with those found by Kim (2021) in other apps (ES = 0.29), even as they were generated when the intervention was delivered for home usage without teacher supervision. In contrast, previous interventions were conducted in schools (Kim et al., 2021). For practitioners, the mean differences for both groups from pretest to posttest displayed in Table 3 indicates that, on average, the children that received the home-based intervention advanced twice more than the advance of the others for number identification, count-sequence, geometric shape identification and spatial sense skills. Future studies may also try to correlate learning analytics extracted by the intervention games with the differences in the GEMMA assessment scores from pretest to posttest. That could lead to a better understanding about why some students benefit more from the intervention than the others.

Regarding set comparison, students of the experimental group had higher pretest scores then those of the control group, probably leaving less room for improvement for some students. One aspect that supports such a possibility is that we observed a reduction among the experimental group in its dispersion that was 4.67 times larger than the reduction of the control group. This reduction in dispersion may indicate that students of the experimental group that had lower scores improved during the intervention, approximating them from those children that had higher scores.

The disparity between students of the same classrooms was exacerbated by the COVID-19 pandemic in many countries. For example, in Europe, standardized assessments revealed significant increases in mathematics learning inequality (Blaskó et al., 2022; Maldonado & De Witte, 2022). In Brazil, school closures reduced early language and mathematics learning ($d = 0.09$) (Bartholo et al., 2022). Children from low-income families attained approximately 48% of the knowledge they would have otherwise (Bartholo et al., 2022).

In the present study, the experimental group students displayed more homogeneous scores after receiving the intervention in all five skills as noted in Table 3. The overall reduction in disparities from pretest to posttest of the experimental group was approximately 9.1 times higher than the reduction observed in the control group. Such large improvement indicates that the intervention contributed to reducing learning inequalities of students from low-income families during school closures.

Compared to benchmarks of other mathematics interventions (Pellegrini et al., 2021), the home-based effects of Escribo Play in count-sequence, geometric shape recognition, and spatial sense skills were more potent than programs offering professional development and traditional or digital curriculum. It also displayed larger effect sizes than other mathematics interventions focusing on first-grade students (Kraft, 2020). The higher gains of Escribo Play intervention may be attributed to its interactive nature, employing evidence-based instructional strategies with engaging and interactive content (i.e., animation and games).

Compared with the cost of other educational interventions (e.g., professional development, tutoring), Escribo Play presents a cost-effective mechanism for decreasing disparities for low-income students. The cost per pupil (USD $50) is much lower than what has been reported in 90% of studies that reviewed other educational interventions ($M$ = $4.752, $SD$ = $9.720) (Kraft, 2020). When compared to the benchmarks (Kraft, 2020), Escribo Play is positioned as easy to scale and provides medium to large learning gains for first-grade students.

While the potential benefits of technology for learning were demonstrated, one aspect to consider is screen time. Detrimental health effects of screen time were found for multiple age groups (Fang et al., 2019; Wang, H. & Lin, 2019). But it is important to mention that screen time can be divided between passive watching content (e.g., video), which is detrimental to child development (Radesky &

Christakis, 2016; Rocha et al., 2021) and interactive screen time, such as playing educational games, which were found to provide beneficial effects in domains such as early literacy (Amorim et al., 2020) and mathematics (Tokac et al., 2019).

### *4.4. Limitations*

There are several limitations with the present study. One limitation of the study is that we do not know the reasons why an important percentage of participants did not take part of the intervention. Due to the early COVID-19 restrictions and stressors affecting the participating teachers' and families, we were not able to conduct interviews or collect survey data remotely. We also do not know why some students who participated did not complete the intervention. Better understanding participation and retention could be the aim for future studies of home-based digital interventions. Another limitation of the study regards the instruction provided to the control group. We did not observe or have data regarding the degree that the children of the control group actually received the expected mathematics instruction specified in the curricula from the emergency remote teaching efforts. If instruction was not provided for some students, the efficacy results of the Escribo Play found in this study may be bigger than they would be if all control students had received proper instruction.

The results may also differ as students were assessed remotely for the pretest and posttest using the GEMMA instrument. We could not control several contextual factors that may have altered the data that were collected. For example, a student may have been helped by older siblings or by the parents. Moreover, perhaps the participants who continued in the program were those whose parents supported them more than the students who did not complete the intervention or who never uploaded the app. Future studies may seek to employ multimodal learning analytics to solve these risks. Employing facial recognition engines that can generate metadata based on the smartphone camera without having to record the images of the students and families may be a feasible path.

## 5. Conclusion

The present study demonstrated the potential of delivering evidence-based, low-cost, mobile-based, early mathematics intervention for students from low-income families at scale in low-income communities. The possibility of deploying low-cost mobile apps that can generate strong learning outcomes when used by the children in their households may be considered by policymakers when designing large-scale supplementary educational efforts. Such interventions may promote equity in learning even when schools are not closed, supporting low-income students to meet learning standards and develop their fullest academic and life potential.

## Credit statement

Americo N. Amorim: Conceptualization, Investigation, Formal analysis, Funding acquisition, Writing, Review & Editing; Lieny Jeon: Conceptualization, Funding acquisition, Writing - review & editing; Yolanda Abel: Conceptualization, Writing - review & editing; Stephen Pape: Writing - review & editing; Emilia X. S. Albuquerque: Investigation, Resources; Monique Soares: Investigation, Resources; Vanessa C. Silva: Investigation, Resources; Danilo Aguiar: Investigation, Resources; José R. Oliveira Neto: Software; Claudia Costin: Investigation; Rodrigo L. Rodrigues: Formal analysis, revision; Mariana de Leon: Conceptualization, Funding acquisition; Carla A. de Paula: Investigation; Jefferson Lopes: Software; Maxsuel S. Silva: Investigation; Maria V. do Nascimento: Investigation; Gabriella A. Patricio: Investigation; Vinícius F. da Silva: Software; Raiane Florentino: Writing - review & editing.

## Data availability

Data can be downloaded from https://doi.org/10.7910/DVN/WIGYFI

## Acknowledgments

This work was funded by Google [Google Cloud COVID-19 research credits], Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq [RHAE Inovação/Bolsa de Produtividade em Desenvolvimento Tecnológico e Extensão Inovadora], Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco – FACEPE [Programa TECNOVA], and Financiadora de Estudos e Projetos – FINEP [TECNOVA/Programa de Aceleração Espaço Finep].

Special thanks to the leadership, teachers, families, and the children of Nova Iguaçu (RJ), Casimiro de Abreu (RJ), and SESC (PE) who participated in this research.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compedu.2023.104759.

## References

Amorim, A. N., Jeon, L., Abel, Y., Felisberto, E. F., Barbosa, L. N., & Dias, N. M. (2020). Using escribo play video games to improve phonological awareness, early reading, and writing in preschool. *Educational Researcher, 49*(3), 188–197.

Bartholo, T. L., Koslinski, M. C., Tymms, P., & Castro, D. L. (2022). *Learning loss and learning inequality during the covid-19 pandemic*. https://doi.org/10.1590/s0104-40362022003003776. Ensaio (Rio De Janeiro, Brazil).

Amorim, A. N., Jeon, L., Abel, Y., Albuquerque, E. X., Soares, M., Silva, V. C., & Neto, J. R. O. (2022). Escribo play learning games can foster early reading and writing for low-income kindergarten children. *Computers & Education, 177*, 104364.

Benavides-Varela, S., Zandonella Callegher, C., Fagiolini, B., Leo, I., Altoè, G., & Lucangeli, D. (2020). Effectiveness of digital-based interventions for children with mathematical learning difficulties: A meta-analysis. *Computers in Education, 157*, Article 103953. https://doi.org/10.1016/j.compedu.2020.103953

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 national survey. *Journal of School Psychology, 72*, 29–48. https://doi.org/10.1016/j.jsp.2018.12.004

Blaskó, Z., Costa, P.d., & Schnepf, S. V. (2022). Learning losses and educational inequalities in europe: Mapping the potential consequences of the COVID-19 crisis. *Journal of European Social Policy, 32*(4), 361–375. https://doi.org/10.1177/09589287221091687

Caregnato, C., Gausmann Pfitscher, R., Raizer, L., & Sfredo Miorando, B. (2019). Desigualdades encadeadas no sistema educacional brasileiro: Estratificações entre os níveis médio e superior da educação. *Práxis Educativa (Ponta Grossa, Paraná, Brazil), 14*(2), 469–486. https://doi.org/10.5212/PraxEduc.v.14n2.004

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal, 45*(2), 443–494. https://doi.org/10.3102/0002831207312908

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. *Acad. Pr.*

Dorouka, P., Papadakis, S., & Kalogiannakis, M. (2020). Tablets and apps for promoting robotics, mathematics, STEM education and literacy in early childhood education. *International Journal of Mobile Learning and Organisation, 14*(2), 255–274. https://doi.org/10.1504/IJMLO.2020.106179

Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: Using nonlinear growth models to estimate student learning capacity. *Educational Researcher, 46*(6), 284–292.

Du, X., Yang, J., Shelton, B. E., Hung, J., & 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

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders, 227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Fang, K., Mu, M., Liu, K., & He, Y. (2019). Screen time and childhood overweight/obesity: A systematic review and meta-analysis. *Child: Care, Health and Development, 45*(5), 744–753. https://doi.org/10.1111/cch.12701

Hawes, Z. C. K., Gilligan-Lee, K. A., & Mix, K. S. (2022). Effects of spatial training on mathematics performance: A meta-analysis. *Developmental Psychology, 58*(1), 112–137. https://doi.org/10.1037/dev0001281

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). *The difference between emergency remote teaching and online learning*.

Kim, J., Gilbert, J., Yu, Q., & Gale, C. (2021). Measures matter: A meta-analysis of the effects of educational apps on preschool to grade 3 children's literacy and math skills. *AERA Open, 7*. https://doi.org/10.1177/23328584211004183, 233285842110041.

Kraft, M. A. (2020). *Interpreting effect sizes of education interventions*. Educational Researcher. https://doi.org/10.3102/0013189X20912798, 0013189X20912798.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588.

Lenhard, W., & Lenhard, A. (2016). *Computation of effect sizes*. https://www.psychometrica.de/effect_size.html.

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K., & Busick, M. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*.

Maldonado, J. E., & De Witte, K. (2022). The effect of school closures on standardised student test outcomes. *British Educational Research Journal, 48*(1), 49–94. https://doi.org/10.1002/berj.3754

Ministério da Educação. (2018). *Base nacional comum curricular*. Ministério da Educação.

Moir, T. (2018). Why is implementation science important for intervention design and evaluation within educational settings? Paper presented at the. *Frontiers in Education, 3*, 61, 2018.

Moro, C., & de Souza, G. (2014). Produção acadêmica brasileira sobre avaliação em educação infantil: Primeiras aproximações. *Estudos Em Avaliação Educacional, 25*(58), 100–125.

Nelson, G., & McMaster, K. L. (2019). The effects of early numeracy interventions for students in preschool and early elementary: A meta-analysis. *Journal of Educational Psychology, 111*(6), 1001–1022. https://doi.org/10.1037/edu0000334

Patro, S., & Sahu, K. K. (2015). *Normalization: A preprocessing stage*. arXiv Preprint arXiv:1503.06462.

Pellegrini, M., Lake, C., Neitzel, A., & Slavin, R. E. (2021). Effective programs in elementary mathematics: A meta-analysis. *AERA Open, 7*. https://doi.org/10.1177/2332858420986211

Platas, L. M., Ketterlin-Geller, L. R., & Sitabkhan, Y. (2016). Using an assessment of early mathematical knowledge and skills to inform policy and practice: Examples from the early grade mathematics assessment. *International Journal of Education in Mathematics, Science and Technology, 4*(3), 163–173.

QEdu. (2020). *Aprendizado dos alunos: Brasil*.

Radesky, J. S., & Christakis, D. A. (2016). Increased screen time: Implications for early childhood development and behavior. *Pediatric Clinics of North America, 63*(5), 827–839. https://doi.org/10.1016/j.pcl.2016.06.006

Rocha, H. A. L., Correia, L. L., Leite, Á. J. M., Machado, M. M. T., Lindsay, A. C., Rocha, Sabrina Gabriele Maia Oliveira, Campos, J. S., Cavalcante E Silva, A., & Sudfeld, C. R. (2021). Screen time and early childhood development in Ceará, Brazil: A population-based study. *BMC Public Health, 21*(1), 2072. https://doi.org/10.1186/s12889-021-12136-2

Russo, J. M., Williford, A. P., Markowitz, A. J., Vitiello, V. E., & Bassok, D. (2019). Examining the validity of a widely-used school readiness assessment: Implications for teachers and early childhood programs. *Early Childhood Research Quarterly, 48*, 14–25. https://doi.org/10.1016/j.ecresq.2019.02.003

Schleicher, A. (2019). *PISA 2018: Insights and interpretations*.

Scull, J. A., & Lo Bianco, J. (2008). Successful engagement in an early literacy intervention. *Journal of Early Childhood Literacy, 8*(2), 123–150.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning, 33*(1). https://doi.org/10.1111/jcal.12172

Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning, 37*(1), 127–141. https://doi.org/10.1111/jcal.12473

Snell, E. K., Wasik, B. A., & Hindman, A. H. (2022). Text to talk: Effects of a home-school vocabulary texting intervention on prekindergarten vocabulary. *Early Childhood Research Quarterly, 60*, 67–79. https://doi.org/10.1016/j.ecresq.2021.12.011

Tokac, U., Novak, E., & Thompson, C. G. (2019). Effects of game-based learning on students' mathematics achievement: A meta-analysis. *Journal of Computer Assisted Learning, 35*(3), 407–420. https://doi.org/10.1111/jcal.12347

Wang, H., & Lin, C. (2019). An investigation into visual complexity and aesthetic preference to facilitate the creation of more appropriate learning analytics systems for children. *Computers in Human Behavior, 92*, 706–715. https://doi.org/10.1016/j.chb.2018.05.032

Yang, W., Liu, H., Chen, N., Xu, P., & Lin, X. (2020). Is early spatial skills training effective? A meta-analysis. *Frontiers in Psychology, 11*, 1938. https://doi.org/10.3389/fpsyg.2020.01938