Abstract—Game-based learning (GBL) is becoming a more widespread style of education; however, there is a lack of research on the benefits of GBL on students’ early literacy skills. This study explores the effectiveness of a computer-adaptive GBL curriculum, Waterford Early Learning (WEL) during the 2016-2017 school year. Kindergarten, first grade, and second grade students took a district-administered literacy assessment at the beginning and end of the school year. Reading assessment scores of kindergarten, first grade, and second grade students who used GBL were compared to scores of students who either did not use GBL or had low usage of WEL during the 2016-2017 school year. Students who used WEL significantly outperformed students who did not use WEL on all end of year literacy scores, and students with high usage of WEL significantly outperformed students with low usage on all end of year literacy scores as well. Students who used WEL also outperformed their control counterparts across demographics. These results indicate that WEL has a potentially positive impact on early literacy skills. More impactful studies concerning incorporating GBL curriculum in conjunction with traditional, in-class literacy instruction are necessary.

Index Terms—technology, game-based learning, literacy

I. INTRODUCTION

The future of education policy is being set by a generation of parents and teachers that understands the utility of serious games. These are parents who grew up with the likes of Reader Rabbit, Math Blaster, The Oregon Trail, and Carmen Sandiego [1]. A recent survey of pre-kindergarten through sixth grade teachers showed that a growing number of educators were using game-derived practices to identify students in need of extra help [2]. There are teachers whose lesson plans for a given year could, to a certain perspective, read like the dust jacket of a fantasy novel, complete with quests to follow, rewards to win, and bosses to overcome [3]. These individuals should be provided with the most accurate information possible to ensure that all students have access to proven and effective curricula.

Game-Based Learning (GBL) refers to adding play-derived elements drawn from game theory to the educational process [4], [5]. A sense of fun and engagement with the material is encouraged with visuals and narrative elements. Students are willing to spend more time with GBL lessons than on traditional classroom curricula [6]. In a GBL lesson, a student is expected to actively participate rather than passively absorb material [7]. An environment that enables exploration is fostered, where students can confront and overcome challenging problems appropriate to their own level of competence. GBL provides students with a sense of achievement and forward progress with meaningful, real-time feedback [7].

Research has shown that GBL interventions can aid in acquiring literacy skills. Children using a GBL curriculum running on a commercial gaming console saw significant gains in high frequency words, active decoding, total reading, and sight words [8]. Students actively participating benefited in a way that students passively absorbing the same material did not. Studies have demonstrated repeatedly, and across multiple languages, that GBL instruments promote strong vocabularies and that participants prefer GBL conditions over traditional alternatives [9], [10]. Serious games have been used to help students with dyslexia acquire foundational skills across multiple continents [11]. GBL literacy instruments, and particularly digital GBL programs, are becoming a ubiquitous presence in the classroom [12].

School budgets are not unlimited. This has been demonstrated recently in the United States by a series of strikes and walk outs over teacher compensation [13]. The decision to use a given GBL instrument should be based on research-based evidence that it will have a tangible positive impact on students’ academic trajectories [14]. While the available body of literature broadly supports the efficacy of GBL [4], [5], this support is not unanimous. Some research highlighted that GBL may best be used as a supplement to, rather than a replacement of, traditional instruction [5]. When implemented with fidelity, digital GBL has been found to be effective for all populations, with specific benefits for young learners [15] and vulnerable populations [16]. However, a recent case study comparing teacher-directed instruction and GBL for acquisition of sight words for children with autism found participants both preferred and saw greater benefit from the teacher-directed condition [17]. This raises the troubling possibility that GBL can fail to engage the most vulnerable students in a meaningful fashion. Some research has found no statistically significant advantage associated with use of GBL beyond pre-existing differences between test sites which were implementing technology and those that were not [2].

Literature stresses the importance of taking valid
Instructional design into account, emphasizing the issue of poorly implemented or designed GBL curricula [6]. Games can fail to engage participants when the focus is too narrow or based on rote memorization of material [18]. It should be noted that the efficacy of digital GBL has been better documented for some groups than for others: A recent literacy survey of GBL studies found only 6% of the sampled literature addressed early learners [12]. Considering the lack of research into GBL’s efficacy with young learners, research is more necessary than ever.

The purpose of the current study was to explore the benefits of a digital GBL instrument for young learners in kindergarten through second grade. It is predicted that students who used GBL software will have better learning outcomes compared to control students.

II. METHODS

A. Participants

This study consisted of kindergarten, first grade, and second grade students (N = 14,307) enrolled in a public school district in Maryland during the 2016-2017 school year.

For kindergarten, the experimental group (n = 967) consisted of students who used WEL for more than 2,000 minutes. The control group (n = 723) consisted of students who used WEL for less than 500 minutes.

For first grade, the experimental group (n = 4,032) consisted of students who used WEL. The control group (n = 1,680) consisted of students who did not use WEL.

For second grade, the experimental group (n = 4,018) consisted of students who used WEL. The control group (n = 2,887) consisted of students who did not use WEL.

B. Materials

1) Waterford Early Learning (WEL)

The program offers a comprehensive, computer-adaptive pre-reading and reading curriculum for pre-kindergarten through second grade students. The software presents a wide range of multimedia-based activities in an adaptive sequence tailored to each student’s initial placement and his or her individual rate of growth throughout the complete reading curriculum.

2) District-administered literacy assessment

The literacy assessment administered by the school district consisted of four substrands, including Known Words, Emergent Behavior, Dictation Sounds, and Dictation Words.

C. Procedure

Kindergarten students were expected to use WEL for fifteen minutes per day, five days per week, and first and second grade students were expected to use WEL for thirty minutes per day, five days per week. Usage was tracked within the program, and total minutes of usage of WEL for the school year per group was calculated. The district assessment was administered at the beginning and at the end of the school year.

III. RESULTS

A. Kindergarten Group Differences using Analysis of Covariance (ANCOVA)

ANCOVAs examining group differences in end of year scores, covarying for beginning of year scores, between the experimental and control groups were conducted (see Fig. 1).

1) Known words

Analysis of Known Words end of year scores, covarying for beginning of year scores, revealed a significant difference between groups, F(1, 1523) = 52.31, p < .01, due to the higher end of year scores made by the experimental students (M = 32.30) than by control students (M = 28.91). Effect size (d = 0.35).

2) Emergent behavior

Analysis of Emergent Behavior end of year scores, covarying for beginning of year scores, revealed a significant difference between groups, F(1, 1519) = 65.94, p < .01, due to the higher end of year scores made by the experimental students (M = 14.96) than by control students (M = 13.99). Effect size (d = 0.38).

3) Dictation sounds

Analysis of Dictation Sounds end of year scores, covarying for beginning of year scores, revealed a significant difference between groups, F(1, 1451) = 51.00, p < .01, due to the higher end of year scores made by the experimental students (M = 13.33) than by control students (M = 11.89). Effect size (d = 0.35).

4) Dictation words

Analysis of Dictation Words end of year scores, covarying for beginning of year scores, revealed a significant difference between groups, F(1, 1455) = 35.75, p < .01, due to the higher end of year scores made by the experimental students (M = 4.51) than by control students (M = 4.00). Effect size (d = 0.29).

B. Kindergarten Group Differences using ANCOVAs by Demographics

1) Known words

Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Known Words end of year scores, covarying for beginning of year scores (see Fig. 2).
a. **Ethnicity**

There was no significant interaction between the effects of ethnicity and WEL on Known Words end of year scores, covarying for beginning of year scores, $F(5, 1512) = 1.26, p = .279$. Simple effects analysis showed that for African American and Hispanic students, students in the experimental group significantly outperformed students in the control group. Caucasian students’ scores in the experimental group were slightly higher than the control group, but the difference was not significant.

b. **Gender**

There was no significant interaction between the effects of gender and WEL on Known Words end of year scores, covarying for beginning of year scores, $F(1, 1521) = 0.01, p = .916$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

2) **Emergent behavior**

Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Emergent Behavior end of year scores, covarying for beginning of year scores (see Fig. 3).

a. **Ethnicity**

There was no significant interaction between the effects of ethnicity and WEL on Emergent Behavior end of year scores, covarying for beginning of year scores, $F(5, 1508) = 1.35, p = .242$. Simple effects analysis showed that for African American and Hispanic students, students in the experimental group significantly outperformed students in the control group. Caucasian students’ scores in the experimental group were slightly higher than the control group, but the difference was not significant.

b. **Gender**

There was no significant interaction between the effects of gender and WEL on Emergent Behavior end of year scores, covarying for beginning of year scores, $F(1, 1517) = 0.16, p = .689$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

3) **Dictation sounds**

Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Dictation Sounds end of year scores, covarying for beginning of year scores (see Fig. 4).

a. **Ethnicity**

There was no significant interaction between the effects of ethnicity and WEL on Dictation Sounds end of year scores, covarying for beginning of year scores, $F(4, 1441) = 1.76, p = .135$. Simple effects analysis showed that for African American and Hispanic students, students in the experimental group significantly outperformed students in the control group. Caucasian students’ scores in the experimental group were slightly higher than the control group, but the difference was not significant.

b. **Gender**

There was no significant interaction between the effects of gender and WEL on Dictation Sounds end of year scores, covarying for beginning of year scores, $F(1, 1449) = 0.00, p = .989$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

4) **Dictation words**

Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Dictation Words end of year scores, covarying for beginning of year scores (see Fig. 5).

a. **Ethnicity**

There was no significant interaction between the effects of ethnicity and WEL on Dictation Words end of year scores, covarying for beginning of year scores, $F(4, 1445) = 0.88, p = .477$. Simple effects analysis showed that for African American and Hispanic students,
students in the experimental group significantly outperformed students in the control group. Caucasian students’ scores in the experimental group were slightly higher than the control group, but the difference was not significant.

Figure 5. Kindergarten end of year dictation words scores by demographics

b. Gender
There was no significant interaction between the effects of gender and WEL on Dictation Words end of year scores, covarying for beginning of year scores, $F(1, 4468) = 1313.21, p < .01$, due to the higher end of year scores made by the experimental students ($M = 48.68$) than by control students ($M = 34.98$). Effect size ($d = 0.85$).

C. First Grade Analysis of Covariance (ANCOVA)
ANCOVAs examining group differences in end of year scores covarying for beginning of year scores, between the experimental and control groups were conducted (see Fig. 6).

1) Known words
Analysis of Known Words end of year scores, covarying for beginning of year scores, revealed a significant difference between groups, $F(1, 4543) = 839.12, p < .01$, due to the higher end of year scores made by the experimental students ($M = 48.68$) than by control students ($M = 34.98$). Effect size ($d = 0.85$).

2) Emergent behavior
Analysis of Emergent Behavior end of year scores, covarying for beginning of year scores, revealed a significant difference between groups, $F(1, 4440) = 171.44, p < .01$, due to the higher end of year scores made by the experimental students ($M = 9.78$) than by control students ($M = 7.51$). Effect size ($d = 0.50$).

D. First Grade Group Differences by Demographics using ANCOVAs
1) Known words
Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Known Words end of year scores, covarying for beginning of year scores (see Fig. 7).

a. Ethnicity
There was no significant interaction between the effects of ethnicity and WEL on Known Words end of year scores, covarying for beginning of year scores, $F(6, 4531) = 0.34, p = .914$. Simple effects analysis showed that for African American, Asian, Caucasian, and Hispanic students, students in the experimental group significantly outperformed students in the control group.

b. Gender
There was no significant interaction between the effects of gender and WEL on Known Words end of year scores, covarying for beginning of year scores, $F(1, 4541) = 2.17, p = .141$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

2) Emergent behavior
Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Emergent Behavior end of year scores, covarying for beginning of year scores (see Fig. 8).
There was a significant interaction between the effects of ethnicity and WEL on Emergent Behavior end of year scores, covarying for beginning of year scores, $F(6, 4428) = 3.90, p < .01$. Simple effects analysis showed that for African American, Asian, Caucasian, and Hispanic students, students in the experimental group significantly outperformed students in the control group.

b. Gender

There was no significant interaction between the effects of gender and WEL on Emergent Behavior end of year scores, covarying for beginning of year scores, $F(1, 4438) = 1.21, p = .272$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

3) Dictation sounds

Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Dictation Sounds end of year scores, covarying for beginning of year scores (see Fig. 9).

a. Ethnicity

There was no significant interaction between the effects of ethnicity and WEL on Dictation Sounds end of year scores, covarying for beginning of year scores, $F(6, 4456) = 1.38, p = .531$. Simple effects analysis showed that for African American, Asian, Caucasian, and Hispanic students, students in the experimental group significantly outperformed students in the control group.

b. Gender

There was no significant interaction between the effects of gender and WEL on Dictation Sounds end of year scores, covarying for beginning of year scores, $F(1, 4473) = 0.50, p = .481$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

4) Dictation Words

Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Dictation Words end of year scores, covarying for beginning of year scores (see Fig. 10).

a. Ethnicity

There was no significant interaction between the effects of ethnicity and WEL on Dictation Words end of year scores, covarying for beginning of year scores, $F(6, 4466) = 1.42, p = .233$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

b. Gender

There was no significant interaction between the effects of gender and WEL on Dictation Words end of year scores, covarying for beginning of year scores, $F(1, 4477) = 0.50, p = .481$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

E. Second Grade Analysis of Covariance (ANCOVA)

ANCOVAs examining group differences in end of year scores, covarying for beginning of year scores, were conducted (see Fig. 11).

1) Dictation Sounds

Analysis of Dictation Sounds end of year scores, covarying for beginning of year scores, revealed a significant difference between groups, $F(1, 5607) = 3636.95, p < .01$, due to the higher end of year scores.
made by experimental students ($M = 61.89$) than by control students ($M = 45.47$). Effect size ($d = 1.37$).

2) Dictation words

Analysis of Dictation Words end of year scores, covarying for beginning of year scores, revealed a significant difference between groups, $F(1, 5624) = 4768.75$, $p < .01$, due to the higher end of year scores made by experimental students ($M = 17.22$) than by control students ($M = 10.08$). Effect size ($d = 1.73$).

F. Second Grade Group Differences by Demographics using ANCOVAs

1) Dictation sounds

Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Dictation Sounds end of year scores, covarying for beginning of year scores (see Fig. 12).

![Figure 12. Second grade end of year dictation sounds scores by demographics.](image)

- **Ethnicity**

There was a significant interaction between the effects of ethnicity and WEL on Dictation Sounds end of year scores, covarying for beginning of year scores, $F(5, 5596) = 6.25$, $p < .01$. Simple effects analysis showed that for African American, Asian, Caucasian, and Hispanic students, students in the experimental group significantly outperformed students in the control group.

- **Gender**

There was a significant interaction between the effects of gender and WEL on Dictation Sounds end of year scores covarying for beginning of year scores, $F(5, 5613) = 8.54$, $p < .01$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

2) Dictation Words

![Figure 13. Second grade end of year dictation words scores by demographics.](image)

Two separate two-way ANCOVAs were conducted to examine the effects of WEL and demographics on Dictation Words end of year scores, covarying for beginning of year scores (see Fig. 13).

- **Ethnicity**

There was a significant interaction between the effects of ethnicity and WEL on Dictation Words end of year scores, covarying for beginning of year scores, $F(5, 5613) = 8.54$, $p < .01$. Simple effects analysis showed that for African American, Asian, Caucasian, and Hispanic students, students in the experimental group significantly outperformed students in the control group.

- **Gender**

There was a significant interaction between the effects of gender and WEL on Dictation Words end of year scores covarying for beginning of year scores, $F(5, 5622) = 15.80$, $p < .01$. Simple effects analysis showed that for females and males, students in the experimental group significantly outperformed students in the control group.

IV. DISCUSSION

The addition of computer-adaptive GBL curriculum as a tool in the classroom has been shown to be generally beneficial to students’ learning [5]. As digital GBL programs become more prominent in schools, the benefits they can have for young learners have not yet been widely studied [12]. The current study supports the hypothesis that digital GBL in a classroom setting can have a positive effect on learning in elementary school students. Students in kindergarten, first grade, and second grade who engaged with WEL as part of their curriculum had significantly higher literacy end of year scores, while covarying for beginning of year scores, compared to students who were exposed only to traditional classroom instruction. These results were observed on all substrands of the literacy test. Scores were consistently higher for the experimental group in both male and female students, as well as for each ethnicity with enough students to be included. Small effect sizes were observed for kindergarten students, and medium to large effect sizes were observed for first grade and second grade students. Results of second grade students had especially large effect sizes for both Dictation Sounds ($d = 1.37$) and Dictation Words ($d = 1.73$). Although this study included sufficiently large samples of students to find statistically significant results, the scope of the study only included a single school district. Certain subgroups, such as students learning English as a second language and those with special needs, were not represented enough to be included in the analysis. Future research on the effects of GBL would benefit from including a wider sample from different, more diverse backgrounds in the hopes of providing more generalizable results. Incorporating a longitudinal design would also provide insight on the long-term effects of GBL for young students, and the impact on students’ academic careers from using GBL at an early age.

REFERENCES

Hawaii International Conference on System Sciences (HICSS), 2014, pp. 82-90.


Haya Shamir was born in Haifa, Israel. Received her PhD from University of Notre Dame in 2002. She has been working at Waterford Research Institute, based in Sandy, Utah, since 2005, currently serves as chief scientist. She has published research in multiple journals throughout her career.

David B. Pocklington was born in Southfield, MI in 1987. Graduated with MS, Experimental Psychology, Rivier University, Nashua, NH, 2014. He currently works as a research assistant for Waterford Research Institute, based in Sandy, UT. Research interests include computer-assisted instruction and assessment.

Kathryn C. Feehan was born in York, Pennsylvania. She graduated with honors from the University of Pittsburgh with a Bachelor of Science in Psychology and Sociology, and she is currently working on her Master of Science in Education at Duquesne University. She currently works at Waterford Research Institute based in Sandy, Utah.

Erik H. Yoder was born in Greenwich, Connecticut, USA. He earned a Bachelor of Arts in psychology and Norwegian from St. Olaf College, Northfield Minnesota, USA, in 2012. He is employed as a research assistant at Waterford Research Institute, based in Sandy, Utah, US.