Gamification and Student Engagement with a Curriculum-Based Measurement System

Yu Yan (✉), Simon Hooper
The Pennsylvania State University, University Park, United States
`yzy122@psu.edu`

Shi Pu
ACT Inc., Iowa City, United States

Abstract—In this study, we employed a random control experiment to evaluate the effectiveness of gamification (e.g. scores, goal, progressive bar, etc.) and initial task difficulty on college student engagement with computer-based assessments. A group of Chinese college students (N=97) were randomly assigned to four groups obtained by crossing the two independent variables: gamification (with or without) and entry level difficulty (low or normal). Students completed several English reading tests (maze tests) for 35 minutes. Student engagement was measured by the average off-task time between two maze tests. The results showed that both gamification and low-difficulty entry level reduced students' off-task time. However, the gamification effect was only significant for male students but not for female students. The study also demonstrated that the maze test can be a potential method to predict general English proficiency with Chinese English language learners.

Keywords—Gamification, engagement, computer-based assessment, curriculum-based measurement, English language learners, task difficulty, gender

1 Introduction

In the late 20th century, notions of effective instruction changed as the theoretical underpinnings of learning expanded from behaviorism to constructivism and social-cultural theory [1]. Our understanding of the role educational assessments likewise shifted from summative assessment to formative assessments, emphasizing that the function of assessment is to provide students with details of their progress to support their learning [2]. Technology plays a critical role in supporting this change by improving efficiency, reducing cost, and assisting in the development of adaptive tasks [3], [4].

This manuscript presents results from the study of a computer-based formative assessment system. This section presents the background and context of the study, describes curriculum-based measurement and gamification, provides a rationale for the current research, and concludes with our research questions.

1.1 Background and context

Echoing the transition from summative to formative assessment, China's Ministry of Education initiated a curriculum reform to encourage computer-based formative assessment in college English teaching and learning. In 2004, the Chinese Ministry of Education (CMOE) began to reform the college English curriculum by implementing two main changes: increasing computer-assisted learning in English to individualize students' learning plans; and, adding formative assessments to the extant summative-heavy assessment system to motivate and inform students about their progress. In response to these reforms, many Chinese universities have developed computer-supported formative assessments to support teaching English—an approach that has been praised by both teachers and students. Formative assessment can assist learning by providing feedback on students' progress. One formative assessment approach involves the use of curriculum-based measurement (CBM). CBM uses reliable, simple, and brief standardized assessments that provide teachers and students with functional performance information that can be used to guide instructional decision making. CBM uses commonplace measurement tasks (e.g., maze tests, reading aloud from a text, and written word sequences), and standardized scoring rubrics to assess each task.

Originally developed to test the effectiveness of an individualized educational plan for special education, CBM is now used widely in elementary and secondary education to monitor students' progress in math, literacy, and language learning in the United States [5]. CBM helps teachers monitor students' learning progress over time and customize instructional interventions for students based on individual performance. Technical adequacy is an important characteristic of CBM [6]. Previous research has demonstrated the reliability and validity of diverse CBM assessments in reading [7], writing [8], and mathematics [9].

Computer administration of CBM can reduce teachers' workload by auto-scoring assessments, improving scoring accuracy, and individualizing feedback. However, to deliver sound and timely information on students' progress, assessments must be valid, reliable, and efficient. Previous research has demonstrated the reliability and validity of using computerized CBM, and teachers have responded favorably to computerized assessment administration [10], [11].

1.2 Curriculum-based measurement and English language learners

Although initially developed to monitor students' math and literacy progress in special education, CBM use has expanded to general education and second language learners [5]. For example, CBM-Reading has been used in the US with English language learners (ELLs) whose native languages are Spanish, Chinese (Mandarin), Japanese, and Arabic, and research has demonstrated that CBM-Reading is sensitive to ELLs' growth in reading skills [12]–[14].

In 2014, 99% percent of Chinese undergraduate students enrolled in regular higher-education institutes were taking English classes [15], [16]. English proficiency plays the role of "gatekeeper" for advanced degrees. Potential graduates are not awarded

bachelor's degrees if their English performance does not meet the national standard. Efficient formative assessment approach is needed for this large population.

However, traditional norm-referenced standardized tests may not be valid for language minority students (i.e., students whose primary or home/native language is not English) due to cultural and linguistic factors [14]. For ELLs, CBM appears to be a promising assessment approach. Thus, it is important, to investigate whether using CBM with Chinese ELLs helps teachers identify students with language learning difficulties. Most studies on CBM and ELLs focus on bilingual students who are learning English in English-speaking countries (for an overview, see Sandberg & Reschly [14]). One recent study was found to investigate the efficacy of using CBM in a non-native English-speaking country. Chung and Espin [17] examined the validity of using CBM with Dutch students who were learning English. They reported alternate-form reliabilities of maze scores ranging from 0.44 to 0.88 as well as correlations between maze scores and English course scores ranging from 0.20 to 0.79. The results suggest that CBM has the potential to be a reliable and valid predictor of ELLs' English proficiency. Because of its small sample size, however, the study can only be considered exploratory. More research is needed to confirm the findings.

1.3 Research on gamification and students' engagement

Encouraging active participation and sustaining motivation while completing computer-based assessments are critical to maintaining the efficacy of formative assessments. Many studies have researched the relationship between students' motivation and their test performance [18], [19]. A literature review of 12 empirical studies indicated that unmotivated students scored more than one-half a standard deviation lower than highly motivated students [20]. Students' motivation also influences the effectiveness of computer-based assessment. For example, research has demonstrated that the extent to which students come to "accept" an assessment system influences their willingness to use CBA [21].

Many strategies have been used to increase students' motivation in online learning and assessment environments. One commonly used approach, gamification, has attracted considerable attention from the education and business sectors [22], [23]. Gamification is the use of game elements in non-gaming contexts to improve user experience and motivation [24]. Results from many gamification studies suggest that gamification has a positive influence on students' motivation and performance. Using gamification in CBM may increase students' participation and help sustain motivation.

1.4 Research questions

The current study addressed the use of a gamified CBM assessment system with Chinese ELLs. We examined the effect of gamification on students' engagement when using a CBM system. In addition, given the benefits of CBM with at-risk school-age students [6], we were interested in whether CBM would help teachers to identify students who experience difficulties learning a second language.

The study included two formal research questions:

RQ1: Do gamification and low-difficulty entry level improve students' engagement when taking assessments?

RQ2: Can the maze test results predict ELLs' English course grades?

2   Method

2.1   Participants

Participants were recruited from a four-year, second-tier regional college in the Sichuan province of southwestern China. The college had an enrollment of approximately 12,000 undergraduate students, of whom 68.4% were female. All freshmen and sophomores were required to take a weekly 90-minute English language course each semester. Students from two English classes taught by one English teacher (N=103) were recruited for the pilot study. However, six students chose not to participate in the study, yielding a final sample of 97 students. The students received extra credit for taking part in the study.

2.2   Materials

The materials included two versions (gamified vs. none-gamified) of a web-based CBM system, known as Avenue: PM. A full description of the software (including the management system, CBM assessments and scoring rules, teacher scoring interfaces, and data performance charts) is presented elsewhere [25]. Students participating in this study only took the maze test in Avenue: PM.

Maze assessments: A maze test is a common CBM assessment used to monitor students' reading progress [26], [27]. In a maze task, students are given a text passage with a blank space for every seventh word, excluding the words in the first sentence. The maze passage includes three choices: one is the correct answer and the other two are distractors. The distractors are taken from a different, randomly selected part of speech rather than the part of speech of the correct answer. For example, if the correct answer were a noun, the two distractors may be verbs, adjectives, determiners, conjunctions, or prepositions. The incorrect choices are randomly generated from a distractors' pool in Avenue: PM. The maze passage is timed to one minute. Students read the passage and use the mouse to select the word for each blank.

Test passages are written at 12 reading levels, with every two levels corresponding to one grade of American students' reading proficiency in primary school. The success criterion (a score from 3–14) is the total score (i.e., correct – incorrect items) required for a student to move forward within a level. Students begin with passages presented at a predetermined reading level (i.e., 1–12), but move up or down levels according to their performance. Each level includes six steps, and students move up or down a step depending on whether they achieve the success criterion for a passage. Reaching step 6 moves the student up a level and falling below step 1 moves the student down a level (in both cases re-starting at step 2). The system uses an algorithm to

select passages in a random order without repetition. After completing a passage, the student receives a score and correct/ incorrect feedback. The system contains 237 passages written by subject-matter experts, and each level includes between 13 and 40 passages.

Two versions of the maze: were developed for this study. Both present students with maze assessments, but they differ in the presence or absence of gaming elements designed to enhance student motivation and engagement.

Version 1: *Maze with gamification features*. Effectively designed game environments create a sense of flow [28], resulting in improved concentration, joy, and involvement [29]. The game elements employed in the maze assessment aim at generating positive emotional experiences by "shift[ing] students' perceptions about the tasks from 'testing' environments to 'practice' or 'gaming' environments" [25].

The gamified version of the software includes four elements (see Figure 1). The first game element is progressive difficulty levels. The test passages become more complex as the levels increase. Students move up or down between levels according to their current reading performances in the system, similar to that which occurs in adaptive assessments. Students receive notices when their current level. The second element is the presence of visually appealing images of animal characters that represent the levels. Low-difficulty levels use characters that are lower on the food chain (e.g., a crustacean or jellyfish) and high-difficulty levels use more advanced animals (e.g., an elephant or lion). Third, a progress bar is presented within each level. The progress bar is similar to the energy graphic that appears in many online games. Students can see where they are in each level and how many more assessments they need to pass before moving on to the next level. The fourth element is immediate feedback on the score. After completing a maze test, students automatically see the correct answers and their total score.
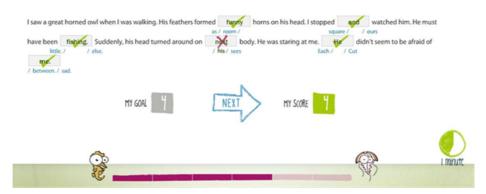

Fig. 1. Maze with gamifications

Version 2: *Maze without gamification features:* The maze test includes embedded game elements. To test the effect of gamification, a non-gamified version of maze was developed and compared with the original maze test with gamification features. Figure 2 includes examples of the non-gamified version. The level notification that is

shown in the gamified system is hidden in the non-gamified version. Table 1 shows the differences between the two versions of maze tests.

Fig. 2. Maze without gamifications

Students using non-gamified maze test do not know their current level, their progress within a level, or their score for each passage. Despite the two versions' different interfaces, the progress mechanism remains the same. Thus, even though students using the non-gamified version are unaware of their progress, they move forward and backward in levels.

Table 1. Table 1. Interface Differences between Two Versions of the Maze

	With gamification	Without gamification
Animal characters indicating levels	Yes	No
Progress bar	Yes	No
Goal for passing each level	Yes	No
Correct score feedback	Yes	No

Entry level difficulty: There are 12 difficulty levels for the maze test. Test passages in successive levels become progressively difficult in both lexical complexity and the success criterion. As described in the previous session, students begin with passages presented at a predetermined level (i.e., 1–12), and move up or down levels according to their performance. Reading passages used in levels 1 – 4 are of low difficulty (i.e., grade equivalent reading level 1 – 2.5), levels 5 – 8 are of normal difficulty (i.e., grade equivalent reading level 3 – 4.5), and levels 9 – 12 are of high difficulty (i.e., grade equivalent reading level 5 – 6.5).

Criterion test: The final exam in the English course was used as the criterion test in this study. The final exam assesses students' English-language ability in different areas, including reading, grammar, vocabulary, and writing. It is a one-hour, paper-based exam taken at the end of the semester. The exam is scored using a range of 0 to 100, with 100 representing the highest possible score.

2.3   Procedure

Students were randomly assigned into four groups obtained by crossing the two independent variables, gamification and entry level difficulty. They used the gamified or non-gamified Avenue: PM system, depending on their group assignment. The students began the maze test in Avenue: PM at either level 3 (low-difficulty entry level) or level 6 (normal entry level). The two different entry levels were suggested by the students' English course teacher. The descriptive statistics covering the students in each treatment group are shown in **Fehler! Verweisquelle konnte nicht gefunden werden.**.

The study was conducted during a 45-minute English class session in a computer lab. Students in the class received a printed handout of the system instructions. The students were given 35 minutes to use the system, but due to an unreliable Internet connection, not all students were able to use the entire 35 minutes for testing.

Table 2. Descriptive Statistics across Treatment Groups

	With gamification	Without gamification	Total
Low-difficulty entry level	$N = 25$	$N = 24$	$N = 49$
Normal entry level	$N = 23$	$N = 25$	$N = 48$
Total	$N = 48$	$N = 49$	$N = 97$

2.4   Data analysis

To examine the impact of gamification and entry level difficulty on students' engagement for research question one, students' average off-task time between maze assessments was used as the dependent variable. The average off-task time was measured by the total off-task time divided by the total number of maze passages completed. Total off-task time and total mazes completed were extracted directly from the database. Higher off-task time indicates lower student engagement. The independent variables included system version (i.e., with or without gamification), entry level difficulty (i.e., low-difficulty or normal), and gender (i.e., male or female).

For research question two, students' final exam scores served as the criterion variable. Students' end level on the maze tests was used as the predicting variable. Pearson correlations were calculated between students' end level on the maze tests and their final score. In addition, differences in maze end level by gender was also examined to assess the validity of the maze test as an indicator of students' English ability.

3   Results

RQ1: Do gamification and difficulty affect students' engagement?

Table 3 reports the mean and standard deviations of off-task time for four treatment groups, and Table 4 reports the mean and standard deviation for the four treatment groups broken down by gender.

Table 3. Students' Off-task Time (in seconds) per Passage by Treatment Groups

Treatment groups	N	Mean	SD
Gamification & low-difficulty entry	25	11.56	6.72
Gamification & normal entry	23	13.60	5.46
No Gamification & low-difficulty entry	24	10.92	6.61
No Gamification & normal entry	25	16.31	6.00
Total	97	13.18	6.47

Table 4. Students' Off-task Time (in seconds) per Passage by Treatment Groups and Gender

Treatment groups	Male				Female	
	N	*Mean*	*SD*	*N*	*Mean*	*SD*
Gamification & low-difficulty entry	9	14.35	10.32	16	9.99	2.86
Gamification & normal entry	8	15.26	6.68	15	12.71	4.70
No Gamification & low-difficulty entry	1	17.64	0.00	23	10.92	6.61
No Gamification & normal entry	3	26.55	8.45	22	14.92	4.17
Total	21	16.60	9.12	76	12.24	5.22

Raw data for the off-task time were not normally distributed, therefore, a log transformation was performed on the off-task time before conducting the subsequent ANOVA analysis. After the log transformation, normality and homogeneity of variance were achieved.

A three-way ($2\times2\times2$) ANOVA was used to examine whether gamification, difficulty, and gender influenced students' engagement as measured by off-task time. The main effect for gamification was significant, [$F(1, 95) = 5.28$, $p < .05$], as was the main effect for low-difficulty entry level [$F(1, 95) = 7.44$, $p < .01$]. The results also showed a significant gender effect, [$F(1, 95) = 9.93$, $p < .01$], with lower off-task time for female students than for male students. The two- and three-way interactions were not significant for all independent variables (all $p > .05$).

To address the concern of "single observation" and the issue of the imbalanced sample in the three-way ANOVA, a follow-up two-sample t-test was conducted to examine the effect of gamification on male and female groups separately. Gamification significantly reduced off-task time for male students ($t = 2.17$, $p < .05$), but not for female students ($t = 0.92$, $p > .05$). The results showed that gamification had a positive effect on reducing the off-task time for a specific gender group.

RQ2: Can maze tests performance predict students' English ability?

To examine the validity of the maze tasks, the correlation between students' end level on the maze tests and their final English score was examined. We also compared the mean differences of maze end level between male and female students. Table 5 reports the Pearson correlation between maze end level and course grade, which ranged from .42 to .48 in the overall sample, within male students, and within female students. All correlations were significant at the .05 level.

Table 5. Correlation between Maze End Level and English Course Exam Score

Gender	N	r	p
Male	21	0.42	0.04
Female	76	0.45	0.04
Total	97	0.48	0.00

The mean difference between gender groups was also examined. For the maze tasks, a statistically significant difference (t = 4.63, p < .001) was found between the two gender groups, with females achieving a higher ending maze level than male students. This result matched the difference in the final course grade between the two gender groups (t = 6.04, p < .001). Table 6 includes mean and standard deviations of students' English test score and maze end level by gender.

Table 6. Students' English Course Exam Score and Maze Tasks End Level by Gender

		English course exam score				Maze tasks' end level			
	N	*Mean*	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
Male	21	48.38	10.43	31	72	4.05	1.687	1	7
Female	75	63.70	10.25	34	83	5.81	1.522	2	9
Total	96	60.38	12.04	31	83	5.43	1.715	1	9

4   Discussion

This study examined the effect of gamification and initial difficulty on Chinese ELLs' engagement in a CBM assessment system. In addition, the study investigated the validity of using maze assessments to measure Chinese ELLs' English proficiency. This section discusses the main findings, implications, and limitations of the study as well as future directions.

Gamification improved engagement by reducing students' off-task time. Using the system with game features (i.e. progress bar, level, and scores) resulted in students spending significantly less off-task time between assessments. This result aligns with previous findings suggesting that gamification can increase students' effort during assessments [30] and improve students' motivation when using a tutoring system [31].

Despite the general positive effect, it is interesting that gamification only reduced male students' off-task time. Two explanations for this finding are offered. First, male students enjoy video games more than do female students [32], [33]; thus, they might be more stimulated and more engaged in the gamified system. Second, levels and points may be more effective for students with lower abilities [30]. In our sample, the female students had higher English proficiency (thus hypothetically, higher English ability) than male students as shown in Table 6, therefore, gamification may have less effect on female students than male students.

However, caution is urged in generalizing these results as the existing literature on the role of gender in gaming environments is mixed. For example, De Jean, Upitis, Koch, and Young [34] found that gender played a key role on learning outcomes and attitude in a gamified learning environment. However, other studies have found gami-

fication to be equally motivating for male and female high school students' in a computer science course [35] or had no gender effect on fifth-graders' math performance and attitude [36]. Recent studies argued that male and female preferred different types of games [37], and females were motivated more by the social aspects of gamification [38]. Therefore, the gender differences in the effectiveness of gamification would be influenced by the design (e.g., the nature of the gamification elements) of the gamification environment.

The current study also demonstrated that low-difficulty entry level improved students' engagement by reducing off-task time while using the system. Students who started at easier levels displayed significantly less off-task time than those starting from normal entry level. Starting with an easier task appears to be a good strategy for improving students' engagement when completing maze tests. Previous research suggested that gaming environments can effectively motivate students by providing "challenging but not overwhelmingly difficult" experiences [39]. When the challenge and difficulty are appropriately balanced, a student will be immersed in a "flow" status [28], in which a person becomes engaged in an activity with deep concentration and enjoyment.

The validity analysis suggests that maze task score is a promising indicator of general English proficiency for Chinese ELLs. First, the correlation between ending maze level and the English course grade is significant for both the overall sample ($r = .48$) and the gender subgroups ($r = .45$ for female; $r = .42$ for male). These findings are consistent with a previous study by Chung and Espin [17], who reported correlations between maze scores and English course grades ranging from 0.19 to 0.79 with Dutch ELLs. In addition, the gender gap in the maze score is consistent with the gender gap in students' English course scores. Gender differences in foreign language learning is a well-acknowledged phenomenon (e.g., [40]). Such differences may be attributable to a difference in brain functions or from the difference in learning strategies[41], [42]. Indeed, Chinese female students usually achieve higher than male students in both English vocabulary and general proficiency tests (e.g., [43]). Therefore, differences in maze performance between male and female groups appear to support the maze as a valid indicator of general English proficiency. In the current study, female students significantly outperformed male students in the English course exam. The analysis indicates that the maze task is sensitive to identifying students with different levels of English proficiency in general.

The current study has three notable strengths. First, the gamified system was not compared to a traditional paper-based assessment or a different computer-based assessment, but to a system identical to the gamified one with the exception of a few gamification features in the interface. Therefore, the differences in students' behaviors were thus due to the gamification effect, not the different format of the test (i.e., paper-based test vs computer-based test). Second, the participants were not children, but college students who may not be easily motivated by gamification elements employed the system (e.g. visually appealing images). Arguably, the effect of the same gamified system on children or adolescents might be larger than the effect found in this study. The third strengthen is that this study used engagement indicators (i.e., off-task time)

derived from students' recorded behaviors in the assessment system, which may be more accurate and objective than data collected from subjective surveys.

Limitations of the study are also noted. The original research question did not consider the gender effect in the design stage; therefore, even though students were randomly assigned to the four treatment groups, the distribution of male and female students was not equal across the four groups. The gamification effect can be more reliable with a balanced design. Furthermore, this study measured students' engagement over a short period of time (35 minutes). Previous research has pointed out that the perceived benefits of gamification decline with use [38]. Therefore, the findings of this study might not be applicable in a long-term study.

Implications for future research are offered. To begin, future study might investigate the effects of voluntary participation[1]. Voluntary participation allows students to enter or leave a game at will. This mechanism can transform a challenging learning process into a more pleasurable experience. This study was only semi-voluntary: students were given 35 minutes during their English class to explore the system. Students did not have an alternative activity available to them. It would be interesting to investigate whether students' behaviors would change without teachers and peers' presence. Second, future studies could consider the interaction effect of the teacher's presence and gamification. While both have the potential to boost student engagement, teachers' presence might diminish the effect of gamification, especially in an educational culture where gaming in class is not a traditional instructional activity but associated with disruptive and dis-engaging behavior that requires teachers' intervention. Third, it would be interesting to investigate the longer-term effects on students' motivation and behaviors while using the Avenue: PM system.

In conclusion, the present study provides support for using the maze task with Chinese ELLs and demonstrates the complexity of designing educational assessment systems with game features. It also opens new directions for further research on using gamification in a CBM system.

5 Acknowledgement

6 References

[1] R. K. Sawyer, "The new science of learning," in The Cambridge handbook of the learning sciences, R. K. Sawyer, Ed. New York: Cambridge University Press, 2006, pp. 1–16.

---

[1] McGonigal [22] defines voluntary participation as one of the four traits of a game, along with a goal, rules, and a feedback system.

[2] National Research Council, Knowing what students know: The science and design of educational assessment. Washington D.C.: National Academies Press, 2001.

[3] J. W. Pellegrino, "Measuring what matters? Technology and the design of assessments that support learning," in Handbook of Design in Educational Technology, R. Luckin, S. Puntambekar, P. Goodyear, B. L. Grabowski, J. Underwood, and N. Winters, Eds. Routledge, 2013, pp. 377–387.

[4] J. W. Pellegrino and E. S. Quellmalz, "Perspectives on the integration of technology and assessment," Journal of Research on Technology in Education, vol. 43, no. 2, pp. 119–134, 2010. https://doi.org/10.1080/15391523.2010.10782565

[5] S. L. Deno, "Curriculum-based measures: Development and perspectives," Assessment for Effective Intervention, vol. 28, no. 3&4, pp. 3–12, 2003. https://doi.org/10.1177/073724 770302800302

[6] S. L. Deno, "Developments in curriculum-based measurement," The Journal of Special Education, vol. 37, no. 3, pp. 184–192, Nov. 2003.

[7] M. M. Wayman, T. Wallace, H. I. Wiley, R. Tichá, and C. A. Espin, "Literature synthesis on curriculum-based measurement in reading," The Journal of Special Education, vol. 41, no. 2, pp. 85–120, Aug. 2007. https://doi.org/10.1177/00224669070410020401

[8] C. A. Espin, J. Shin, S. L. Deno, S. Skare, S. Robinson, and B. Benner, "Identifying indicators of written expression proficiency for middle school students," The Journal of Special Education, vol. 34, no. 3, pp. 140–153, Oct. 2000. https://doi.org/10.1177/00224 6690003400303

[9] P. M. Stecker, L. S. Fuchs, and D. Fuchs, "Using curriculum-based measurement to improve student achievement: Review of research," Psychology in the Schools, vol. 42, no. 8, pp. 795–819, 2005. https://doi.org/10.1002/pits.20113

[10] N. Förster and E. Souvignier, "Curriculum-based measurement: Developing a computer-based assessment instrument for monitoring student reading progress on multiple indicators," Learning Disabilities: A Contemporary Journal, vol. 9, no. 2, pp. 65–88, 2011.

[11] M. Tsuei, "A web-based curriculum-based measurement system for class-wide ongoing assessment," Journal of Computer Assisted Learning, vol. 24, no. 1, pp. 47–60, May 2008. https://doi.org/10.1111/j.1365-2729.2007.00242.x

[12] R. Domínguez de Ramírez and E. S. Shapiro, "Curriculum-based measurement and the evaluation of reading skills of Spanish-speaking English language learners in bilingual education classrooms," School Psychology Review, vol. 35, no. 3, pp. 356–369, 2006.

[13] E. Farmer, "Examining predictive validity and rates of growth in curriculum-based measurement with English language learners in the intermediate grades," Loyola University Chicago, 2013.

[14] K. L. Sandberg and A. L. Reschly, "English learners: Challenges in assessment and the promise of curriculum-based measurement," Remedial and Special Education, vol. 32, no. 2, pp. 144–154, 2011. https://doi.org/10.1177/0741932510361260

[15] National Bureau of Statistics of China, China statistical yearbook. China Statistics Press, 2014.

[16] Q. Wang, "The national curriculum changes and their effects on English language teaching in the People's Republic of China," in International Handbook of English Language Teaching SE - 8, vol. 15, J. Cummins and C. Davison, Eds. Springer US, 2007, pp. 87–105. https://doi.org/10.1007/978-0-387-46301-8_8

[17] S. Chung and C. A. Espin, "CBM progress monitoring in foreign language learning for secondary school students: Technical adequacy of different measures and scoring procedures," Assessment for Effective Intervention, vol. 38, no. 4, pp. 236–248, Jun. 2013. https://doi.org/10.1177/1534508413489723

[18] J. Baumert and A. Demmrich, "Test motivation in the assessment of student skills: The effects of incentives on motivation and performance," European Journal of Psychology of Education, vol. 16, no. 3, pp. 441–462, 2001. https://doi.org/10.1007/bf03173192

[19] L. F. Wolf and J. K. Smith, "The consequence of consequence: Motivation, anxiety, and test performance," Applied Measurement in Education, vol. 8, no. 3, pp. 227–242, 1995. https://doi.org/10.1207/s15324818ame0803_3

[20] S. L. Wise and C. E. DeMars, "Low examinee effort in low-stakes assessment: Problems and potential solutions," Educational Assessment, vol. 10, no. 1, pp. 1–17, 2005. https://doi.org/10.1207/s15326977ea1001_1

[21] V. Terzis and A. A. Economides, "The acceptance and use of computer based assessment," Computers & Education, vol. 56, no. 4, pp. 1032–1044, May 2011. https://doi.org/10.1016/j.compedu.2010.11.017

[22] J. McGonigal, Reality is broken: Why games make us better and how they can change the world. Penguin, 2011.

[23] T. Reiners and L. C. Wood, Eds., Gamification in Education and Business, 1st ed. Springer International Publishing, 2015.

[24] S. Deterding, M. Sicart, L. Nacke, K. O'Hara, and D. Dixon, "Gamification: Using game design elements in non-gaming contexts," in CHI '11 Extended Abstracts on Human Factors in Computing Systems, 2011, pp. 2425–2428. https://doi.org/10.1145/1979742.1979575

[25] S. Hooper, C. Miller, and S. Rose, "Considering the design of an electronic progress-monitoring system," in Handbook on Design in Educational Computing, R. Luckin, P. Goodyear, B. Grabowski, S. Puntambeker, J. Underwood, and N. Winters, Eds. Routledge, 2013. https://doi.org/10.4324/9780203075227.ch22

[26] J. Shin, S. L. Deno, and C. A. Espin, "Technical adequacy of the maze task for curriculum-based measurement of reading growth," The Journal of Special Education, vol. 34, no. 3, pp. 164–172, Oct. 2000. https://doi.org/10.1177/002246690003400305

[27] R. Tichá, C. A. Espin, and M. M. Wayman, "Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading-aloud and maze-selection measures," Learning Disabilities Research & Practice, vol. 24, no. 3, pp. 132–142, Aug. 2009. https://doi.org/10.1111/j.1540-5826.2009.00287.x

[28] M. Csikszentmihalyi, Flow: The psychology of optimal experience. 1990.

[29] D. Johnson and J. Wiles, "Effective affective user interface design in games," Ergonomics, vol. 46, no. 13–14, pp. 1332–1345, 2003. https://doi.org/10.1080/00140130310001610865

[30] Y. Attali and M. Arieli-Attali, "Gamification in assessment: Do points affect test performance?," Computers & Education, vol. 83, pp. 57–63, 2015. https://doi.org/10.1016/j.compedu.2014.12.012

[31] G. T. Jackson and D. S. McNamara, "Motivation and performance in a game-based intelligent tutoring system," Journal of Educational Psychology, vol. 105, no. 4, pp. 1036–1049, 2013. https://doi.org/10.1037/a0032580

[32] C. Chou and M. J. Tsai, "Gender differences in Taiwan high school students' computer game playing," Computers in Human Behavior, vol. 23, no. 1, pp. 812–824, 2007. https://doi.org/10.1016/j.chb.2004.11.011

[33] T. Karakus, Y. Inal, and K. Cagiltay, "A descriptive study of Turkish high school students' game-playing characteristics and their considerations concerning the effects of games," Computers in Human Behavior, vol. 24, no. 6, pp. 2520–2529, 2008. https://doi.org/10.1016/j.chb.2008.03.011

[34] J. De Jean, R. Upitis, C. Koch, and J. Young, "The Story of Phoenix Quest: How girls respond to a prototype language and mathematics computer game," Gender and Education, vol. 11, no. 2, pp. 207–223, 1999. https://doi.org/10.1080/09540259920708

[35] M. Papastergiou, "Digital game-based learning in high school Computer Science education: Impact on educational effectiveness and student motivation," Computers & Education, vol. 52, pp. 1–12, 2009. https://doi.org/10.1016/j.compedu.2008.06.004

[36] F. Ke and B. Grabowski, "Gameplaying for maths learning: Cooperative or not?," British Journal of Educational Technology, vol. 38, no. 2, pp. 249–259, 2007. https://doi.org/10.1111/j.1467-8535.2006.00593.x

[37] T. Hartmann and C. Klimmt, "Gender and computer games: Exploring females' dislikes," Journal of Computer-Mediated Communication, vol. 11, no. 4, pp. 910–931, 2006. https://doi.org/10.1111/j.1083-6101.2006.00301.x

[38] J. Koivisto and J. Hamari, "Demographic differences in perceived benefits from gamification," Computers in Human Behavior, vol. 35, pp. 179–188, 2014. https://doi.org/10.1016/j.chb.2014.03.007

[39] R. M. Ryan, C. S. Rigby, and A. Przybylski, "The motivational pull of video games: A self-determination theory approach," Motivation and Emotion, vol. 30, no. 4, pp. 347–363, 2006. https://doi.org/10.1007/s11031-006-9051-8

[40] J. Sunderland, "Issues of language and gender in second and foreign language education," Language Teaching, vol. 33, no. 04, pp. 203–223, 2000. https://doi.org/10.1017/s0261444800015688

[41] Y. Aida, "Examination of Horwitz, Horwitz, and Cope's construct of foreign language anxiety," The Modern Language Journal, vol. 78, no. 2, pp. 155–168, 1994. https://doi.org/10.1111/j.1540-4781.1994.tb02026.x

[42] J. M. Green and R. Oxford, "A closer look at learning strategies, L2 proficiency, and gender," TESOL quarterly, vol. 29, no. 2, pp. 261–297, 1995. https://doi.org/10.2307/3587625

[43] Y. Gu, "Gender, academic major, and vocabulary learning strategies of Chinese EFL learners," RELC Journal, vol. 33, no. 1, pp. 35–54, 2002. https://doi.org/10.1177/003368820203300102

7 Authors

Yu Yan is a Ph.D. candidate in Learning, Design, and Technology at The Pennsylvania State University, 314 Keller Building, University Park, PA 16802, United States. Her research interests include gamification, student engagement, learning analytics, educational data mining, cognitive load, EEG and flipped classroom.

Simon Hooper is Professor, Learning, Design, and Technology at The Pennsylvania State University, where he designs, develops, and examines the impact of technology to support teaching and learning. His current research interests include progress monitoring, interaction design, learning analytics, and data visualization.

Shi Pu is a data scientist in the AI and Machine Learning team at ACT, Inc, 500 Act Drive, Iowa City, Iowa 52243, United States. He received his Ph.D. degree in Higher Education at The Pennsylvania State University, United States.