Developing a Rasch-Based Scale for Vocabulary Game Leveling

Tzemin Chung[1,*] Mohd Zali Mohd Nor[2] Richard Yan[1] and Peing Ling Loo[1]

[1] CommonTown Pte Ltd., Singapore 408868
[2] Persatuan Rasch Malaysia, Kuala Lumpur, Malaysia
*tzemin_chung@commontown.com

Abstract. Vocabulary acquisition is fundamental to language learning, and educational games provide an engaging way to support vocabulary development. However, designing game challenges that appropriately match learners' abilities remains a significant challenge. This study explores the use of the Rasch measurement model to develop a scale for vocabulary game leveling, enabling game challenges to be calibrated according to their difficulty and aligned with learners' abilities. Gameplay data were collected from 700 students in Singapore and Turkey across five English vocabulary games. Rasch analysis was used to calibrate challenge difficulty, estimate learner ability, evaluate item fit, examine dimensionality, and investigate the alignment between game challenges and student abilities. The results established a meaningful hierarchy of challenge difficulties and showed that the games largely measured a common underlying construct of English language ability. The Wright Map indicated reasonable overall targeting while identifying opportunities to improve the distribution of challenge difficulty. Misfitting challenges were also identified for refinement to enhance measurement quality and curriculum alignment. The findings demonstrate that Rasch-based game leveling provides a practical measurement framework for educational game design and vocabulary practice. The resulting scale offers a systematic approach for matching challenge difficulty to learner ability while informing future game development and adaptive learning applications. Future research may further examine measurement equivalence across different learner populations and expand the range of calibrated game challenges.

Keywords: Rasch model, educational games, game leveling, adaptive learning, measurement scale, English vocabulary learning.

1 Introduction

Vocabulary is the basis of reading comprehension. As students' vocabulary sizes expand, their reading abilities improve. They are also more likely to integrate new knowledge with their existing background knowledge, facilitating improved

© The Author(s) 2026
C. Y. Lye et al. (eds.), *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS 2025)*, Atlantis Highlights in Social Sciences, Education and Humanities 56,
https://doi.org/10.2991/978-94-6239-834-4_20

comprehension. In addition, they tend to read more often, which further contributes to vocabulary growth [1].

Vocabulary should be introduced both directly and indirectly. Direct teaching refers to explicit instruction of pre-selected words before reading a book or participating in class activities, while indirect teaching includes activities that encourage students to build vocabulary through authentic language experiences [2]. As students need to encounter newly learned words multiple times, preferably more than ten times, to acquire them effectively [3, 4], direct instruction alone is insufficient. Students require opportunities to experience meaningful language input beyond the classroom [5].

In addition, merely memorising words from textbooks is not enough to support long-term vocabulary development. Classroom teachers have therefore explored the use of computers to facilitate vocabulary instruction and have found them to be effective learning tools [6, 2]. In particular, educational games have been identified as one of the most effective instructional approaches, including for English second-language learning [7].

According to Self-Determination Theory [8], motivation is driven by the fulfilment of three basic psychological needs: competence, autonomy, and relatedness. Educational games can stimulate student interest in achieving learning goals by fostering a sense of competence through challenges and rewards [1]. They also support autonomy by allowing students to take greater control of their own learning [9] and promote relatedness by encouraging participation, even among shy students [10]. As a result, games help sustain student interest in otherwise routine learning tasks and maintain learning effort over time [9].

Beyond motivation, educational games offer several additional learning benefits. They can be adapted to accommodate different learner needs [11], provide opportunities for repeated review of previously learned vocabulary [9], and have been shown to improve student achievement by an average of 20 percentile points when incorporated into classroom teaching [12, 13]. Given these advantages, educational games provide an effective means of supporting vocabulary learning beyond curriculum time.

As games are powerful learning aids for vocabulary acquisition, this study developed a suite of interactive vocabulary games to help students review high-frequency words drawn from their textbooks. To ensure that game challenges appropriately match learners' abilities, the study applies the Rasch measurement model to develop a Rasch-based scale for vocabulary game leveling. Using gameplay data collected from young learners in Singapore and Turkey, the study calibrates the difficulty of vocabulary game challenges, estimates learner abilities, and evaluates the measurement properties of the resulting scale. The findings demonstrate how Rasch analysis can support educational game design by providing a practical and evidence-based approach to game leveling while contributing to more effective vocabulary learning experiences.

2 Method

2.1 Participants

This study investigated the use of the Rasch measurement model to develop a game-leveling scale for English vocabulary learning. Gameplay data were collected from 700 students aged 5 to 13 years, ranging from Kindergarten 1 (K1) to Secondary 1 (Sec 1, equivalent to Grade 7), who used a suite of interactive English vocabulary games as part of their normal learning activities. The participants included 82 students from Singapore and 618 students from Turkey and represented a wide range of English language abilities, providing a diverse sample for calibrating game challenge difficulty and estimating learner ability. The distribution of participants by grade level is presented in Table 1.

Table 1. Distribution of participants by educational level.

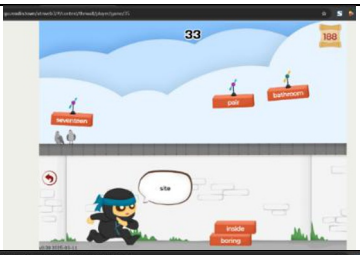
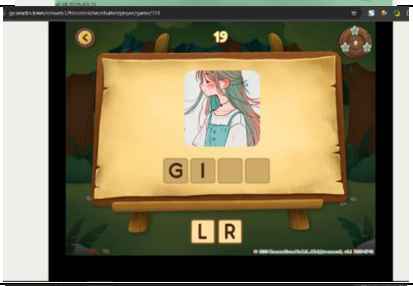
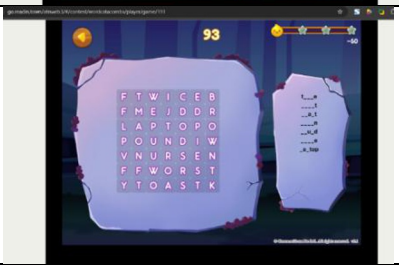
Educational Level	Number of students
Kindergarten 1	17
Kindergarten 2	2
Primary 1	116
Primary 2	148
Primary 3	159
Primary 4	132
Primary 5	84
Primary 6	27
Secondary 1	15
Total	700

2.2 Materials

The study utilised five interactive English vocabulary games designed to support English vocabulary learning through engaging and varied learning activities (Table 2). Each game comprised multiple challenges that targeted different aspects of vocabulary development, including word recognition, oral fluency, word generation, semantic processing, and visual search.

The game challenges were categorised into three difficulty levels based on the Common European Framework of Reference for Languages (CEFR): Easy (Level A1, approximately Grades 1–3), Medium (Level A2, approximately Grades 4–6), and Difficult (Level B1, approximately Grades 7–9). Across the five games, a total of 1,465 individual game challenges were developed to provide learners with progressive vocabulary practice.

Table 2. Characteristics of the five vocabulary games.

Name	Game	Objective	Main Skill
The Wall (TWCH)		Read aloud the falling words.	Word recognition
Quick Speak (QSCH)		Read the highlighted sentence aloud.	Oral fluency; Sentence structure
Word Safari (WSCH)		Create words using the given hints.	Semantic processing; Word generation
Word Finder (WFCH)		Generate words from the given letters.	Word generation; Visual search
Word Catacombs (WCCH)		Find words using the provided clues.	Semantic processing; Visual search; Problem solving

The five games were designed to provide vocabulary practice through a variety of challenge types while supporting different aspects of English language learning. In this study, the individual game challenges served as the measurement units for Rasch calibration, enabling challenge difficulty and learner ability to be estimated on a common measurement scale.

2.3 Procedure

Students accessed the vocabulary games at their own convenience and freely selected from the available challenges. Most games introduced nine new challenges each week, comprising three easy, three medium, and three difficult activities. When a student attempted the same challenge multiple times, only the highest performance score was retained for analysis.

During the study period, the five games provided a total of 1,465 game challenges. Of these, 633 challenges were attempted, generating 5,547 gameplay responses for subsequent Rasch analysis.

2.4 Data Analysis

The gameplay data were analysed using the Rasch Partial Credit Model (PCM) implemented in Winsteps (Version 5.9.0) [14]. The PCM was selected because individual game challenges could receive partial scores rather than simple dichotomous correct or incorrect responses. In this study, the students were treated as persons, while the individual game challenges were treated as items.

To improve the stability of the measurement model, only game challenges attempted by at least ten [15] students were included in the final analysis. This filtering process reduced the dataset from 633 attempted challenges to 161 calibrated game challenges and from 700 participants to 668 students for Rasch analysis.

The Rasch analysis was conducted to estimate learner ability and challenge difficulty on a common measurement scale. Item and person fit statistics were examined to evaluate the consistency of observed responses with model expectations. Reliability and separation indices were used to assess the ability of the game challenges to distinguish between different levels of learner ability and the stability of the challenge hierarchy.

A Wright Map was constructed to examine the targeting of game challenges to learner abilities and to identify gaps in challenge difficulty across the measurement scale. Dimensionality analysis was also performed to investigate whether the game challenges measured a common underlying construct of English vocabulary ability. The findings from these analyses provide evidence for the development and evaluation of a Rasch-based game-leveling scale for vocabulary learning.

3 Results

3.1 Summary Statistics

Table 3 presents the summary statistics from the Rasch analysis, providing an overview of model fit, measurement precision, and reliability for both persons (students) and items (game challenges).

The measure (logit) represents the average ability of persons and the average difficulty of items. The person measure of –0.26 logits indicates that the game challenges were, on average, slightly easier than the students' abilities, while the item measure was centred at 0.00 logits, as required by the Rasch model.

The standard error (SE) reflects the uncertainty associated with the estimated measures. The person SE of 0.59 logits is relatively high, indicating greater uncertainty in estimating student ability. In contrast, the item SE of 0.23 logits is lower, suggesting more stable estimates of game challenge difficulty. According to [16], standard errors below 0.30 logits are generally considered acceptable, whereas values above 0.50 logits indicate greater measurement uncertainty.

The infit mean square (MNSQ) indicates how well observed responses conform to Rasch model expectations. Values between 0.5 and 1.5 are generally considered acceptable [17]. The person infit MNSQ of 0.71 and the item infit MNSQ of 0.76 both indicate slight overfit, suggesting that responses are more predictable than expected but remain within the acceptable range.

The infit ZSTD provides a standardized measure of fit, with values between –2 and +2 considered acceptable [18]. The person infit ZSTD of –0.61 and the item infit ZSTD of –1.03 both fall within this range. Their negative values indicate that the observed responses are more predictable than expected, consistent with the slight overfit indicated by the infit MNSQ statistics.

The outfit MNSQ is more sensitive to unexpected outlying responses and follows the same acceptable range of 0.5 to 1.5 [17]. The person outfit MNSQ of 0.74 and the item outfit MNSQ of 0.79 both suggest slight overfit while remaining within acceptable limits.

Similarly, outfit ZSTD values between –2 and +2 are considered acceptable [18]. The person outfit ZSTD of –0.57 and the item outfit ZSTD of –0.92 both fall within this range. These negative values indicate lower-than-expected response variability and provide further evidence that the data fit the Rasch model satisfactorily despite a slight tendency towards overfit.

The person and item separation indices indicate the extent to which the measurement system distinguishes between different levels of learner ability and game challenge

difficulty. According to Linacre [19], separation values below 1.5 indicate limited discrimination, whereas values above 2.0 suggest good differentiation. In this study, the person separation index of 0.53 indicates that the current set of game challenges had limited ability to distinguish between students of different vocabulary abilities. In contrast, the item separation index of 5.36 demonstrates that the game challenges were well distributed across a broad range of difficulty levels, supporting the development of a meaningful hierarchy of game challenge difficulty.

The person and item reliability coefficients support these findings. Person reliability reflects the consistency with which the instrument distinguishes between learners of different abilities, whereas item reliability indicates the stability of the estimated game challenge difficulties. According to Linacre [19], reliability values above 0.80 are generally considered strong, while values below 0.50 indicate weaker measurement consistency. The person reliability coefficient of 0.22 suggests limited consistency in differentiating between learner abilities, whereas the item reliability coefficient of 0.97 indicates a highly stable hierarchy of game challenge difficulty. These findings suggest that, while the game challenges were successfully calibrated across a wide range of difficulty levels, further refinement may improve the instrument's ability to distinguish between learners with different vocabulary abilities.

Table 3. Summary statistics for the Rasch analysis.

	Measure (logits)	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Separation	Reliability
Person	-0.26	0.59	0.71	-0.61	0.74	-0.57	0.53	0.22
Item	0.00	0.23	0.76	-1.03	0.79	-0.92	5.36	0.97

The low person separation (0.53) and reliability (0.22) in this Rasch analysis are primarily due to insufficient person data and low variability in student performance. Several factors contributed to these results.

1. Too few responses per student.
   Each student responded to an average of only eight game challenges, which is insufficient for stable ability estimation. Many students attempted only a small subset of the 1,465 available game challenges, resulting in sparse response patterns with many missing observations. Since only 633 of the 1,465 challenges were attempted, there is limited data for accurately estimating each student's ability. According to Linacre [16], Rasch analysis generally requires a minimum of 50 responses per person for stable person estimation. Consequently, the low number of responses per student contributed to the low person separation and reliability values.

2. Limited Variation in Student Performance.
   Since the participants ranged from kindergarten (5 years old) to grade 7 (13 years old), most were not yet ready to engage with the higher-level challenges. As a result, responses clustered around the easier items, reducing the range of observed abilities and weakening person separation.

3. Gameplay design and scoring.
   In this dataset, students can attempt challenges multiple times, with only their best attempt being recorded. As a result, variation within each challenge is naturally low, as students may continue playing until they achieve a high score or even full marks. This differs from traditional test-based assessments, where responses are typically recorded in a single sitting, leading to greater variability [20].

These factors help explain the low person separation and reliability values observed in this study. In contrast, the high item separation and reliability indicate that the game challenges themselves were successfully calibrated across a broad range of difficulty levels. This issue becomes more apparent in the Wright Map, which visually displays the distribution of learner abilities relative to game challenge difficulties. The following section examines these patterns in greater detail.

3.2 Item Location and Person Measures

The Wright Map (Figure 1) provides a visual representation of the relationship between student abilities and game challenge difficulties on the same logit scale. Student abilities were concentrated mainly between -2 and +1 logits, whereas challenge difficulties covered a much broader range. The difference between the average person and item measures was relatively small (0.26 logits), indicating reasonable overall targeting of the game challenges to the participating students.

However, the distribution also revealed gaps across the ability continuum. Several higher-difficulty challenges were beyond the observed ability range of most students, while comparatively few challenges targeted the central ability range where many students were located. This pattern suggests opportunities to improve the balance of challenge difficulty and enhance the precision of measurement.

The Wright Map further showed variation across the English language games, with more difficult activities occupying the upper end of the scale and easier activities located toward the lower end. This spread indicates that the game-leveling system captures a broad range of challenge difficulties while highlighting areas where additional content could strengthen coverage of student abilities.

Overall, the Wright Map supports the summary statistics and suggests that future game development should include more challenges targeted at the ability levels represented in the current sample, thereby improving both measurement precision and the educational value of the adaptive game-leveling system.

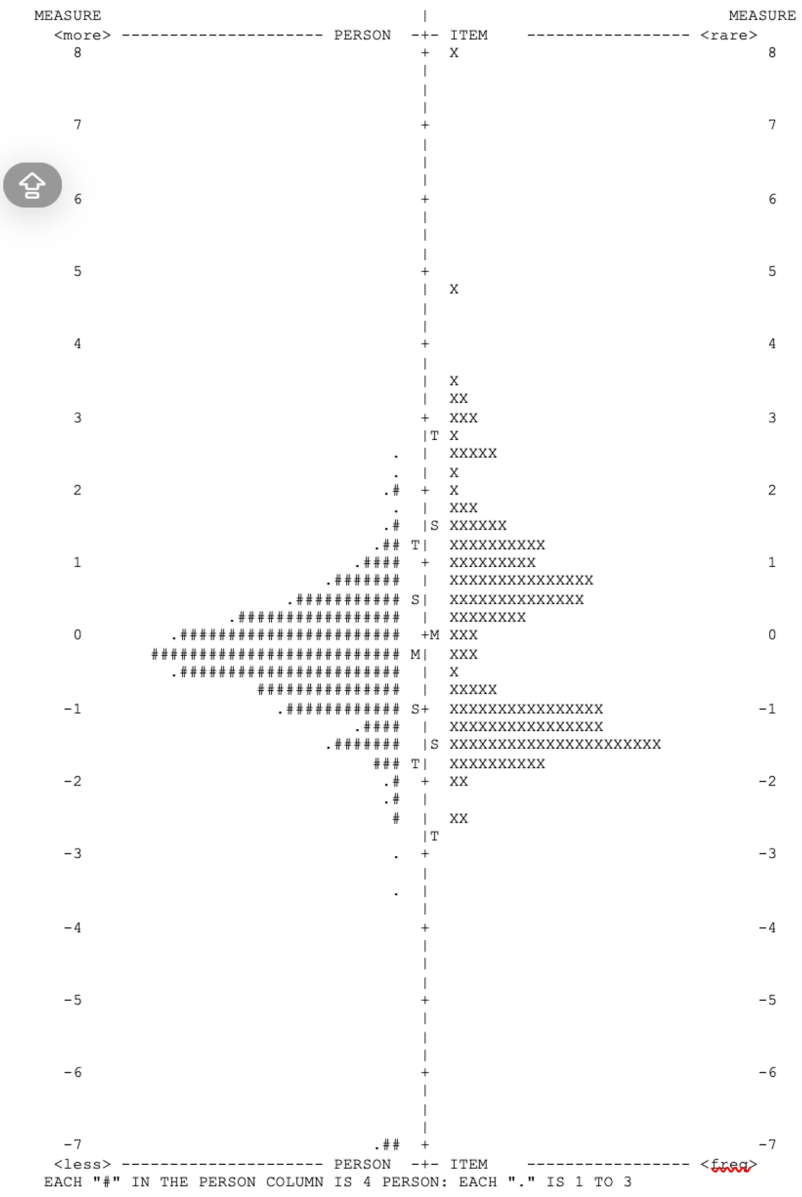


Fig. 1. The Wright Map

3.3 Dimensionality

A dimensionality analysis, conducted using Winsteps, evaluated 161 items across five games (TW, QS, WC, WS, WF) designed for students to practice English, assessing whether they measure a single construct: English language ability. The primary dimension accounted for 75% of the total variance (481.1880/641.1880) (Table 3), exceeding the 40% threshold for unidimensionality in language assessments [14], indicating that the games predominantly measure English language ability.

The first contrast, representing a potential secondary dimension, yielded an eigenvalue of 5.81. While an eigenvalue greater than 2 can suggest multidimensionality, this contrast accounted for only 0.9% of the total variance, which is well below the commonly used 5% threshold. This suggests that any secondary dimension is likely weak and not substantial [14].

Table 4. Explained and residual variance for the dimensionality analysis.

Variance	Eigenvalue	Observed	Expected
Total raw variance in observations	641.2	100.0%	100.0%
Raw variance explained by measures	481.2	75.0%	74.4%
Raw variance explained by persons	129.0	20.1%	20.0%
Raw variance explained by items	352.1	54.9%	54.5%
Raw unexplained variance (total)	160.0	25.0%	
Unexplained variance in 1st contrast	5.8	.9%	
Unexplained variance in 2nd contrast	4.3	.7%	

To explore this secondary dimension, the analysis grouped items into three clusters based on their standardized residual correlations: Cluster 1, comprising five TWCH items and one WFCH item (positive loadings 0.34–0.73), emphasized lexical processing skills, such as word recognition and generation; Cluster 3, with 12 WCCH and four WSCH items (negative loadings -0.70 to -0.24), focused on visual search and problem-solving abilities; and Cluster 2, containing 139 items with near-zero loadings, contributed minimally to the contrast. High disattenuated correlations between clusters (0.8196–1.0000) further supported unidimensionality, as these values, close to 1.0, indicate that the clusters measure the same underlying construct after adjusting for measurement error.

Consequently, the game scores can be aggregated into a single measure of English language ability without requiring separate scoring dimensions. The minor skill variations identified reflect task variation [18], such as lexical processing versus problem-solving, provide opportunities to refine game design and enhance English language practice.

For instance, game design can be improved by adjusting difficulty levels or incorporating scaffolds, such as adding hints for WFCH, which students may find challenging due to its focus on creative word formation (e.g., starting with shorter

words or providing example solutions). Additionally, educators can analyze performance patterns to offer targeted feedback: students struggling with TWCH/QSCH may benefit from support in reading fluency or vocabulary recognition (e.g., practice with word lists or sentence reading exercises), those facing difficulties with WCCH/WSCH could improve through strategies enhancing problem-solving or visual search (e.g., exercises for pattern recognition or structured word-building tasks), and students excelling in TWCH but struggling in WCCH might be strong in reading fluency but need support in problem-solving. These adjustments and interventions can optimize the games' effectiveness for English language learning by addressing skill-specific strengths and weaknesses.

3.4 Item Analysis

Item analysis was conducted to identify problematic game challenges that may require modification. Items are considered problematic if their infit or outfit values exceed 1.5. An outfit value greater than 1.5 suggests that students performed well due to luck or random guessing, while an infit value greater than 1.5 indicates inconsistencies in student responses, potentially signaling a flaw in the item. In this case, 24 items have outfit mean square (MS) values greater than 1.5, and 23 items have infit mean square (MNSQ) values greater than 1.5. Notably, the 23 items flagged by infit overlap with those flagged by outfit.

In this study we focused mainly on the high infit and mean square values. High infit value for an item indicates that it is not discriminating well for the people on whom the items are targeted. It might be too difficult or too easy for most students, or it could be confusing or misaligned with the skills or knowledge (e.g., CEFR levelling) the curriculum is meant to assess [22]. High outfit mean square values can be due to random guessing [22].

Low infit and outfit values, often due to item redundancy, do not significantly impact the effectiveness of our measurement scale [22].

A total of 23 misfitting challenges were identified. Most of these challenges contain off-level words, and a few include possessive nouns that students struggled to pronounce, indicating a misalignment with the curriculum. Below is our analysis of four representative examples, which illustrate the range of issues we encountered.

Table 5. Sample misfitting game challenges.

Challenge	Logit	Infit	Outfit	Example	Issue and Action
WSCH065La	6.85	∞	∞	very wave (B1) see its seven	Wave is a high-level word. Use a more familiar word.

Challenge	Logit	Infit	Outfit	Example	Issue and Action
QSCH003La	-1.00	3.53	4.00	I am from England. She is from Italy. They are from France.	Country names are off list (not leveled). Decide whether to retain or remove them.
QSCH005Ma	0.63	1.90	2.35	The blue cell phone is my brother's.	Possessive nouns (‘s) are difficult for beginners. Introduce them at higher levels.
TWCH125La	0.43	1.75	1.77	app, horse, game, meat, …	"App" is beyond the target vocabulary list. Replace it.

4 Limitations

Several limitations should be considered when interpreting the findings.

First, although 700 students participated, each student completed only a small number of challenges, averaging about eight responses. This limited the precision of person ability estimates and contributed to the low person separation and reliability.

Second, only the highest score for each challenge was retained. Unlike conventional assessments, repeated attempts allow students to improve their performance, reducing response variability and limiting the Rasch model's ability to distinguish between learners.

Third, the participant sample covered a relatively narrow age and ability range, with many higher-level challenges receiving few or no attempts. This reduced the effectiveness of the model in targeting the full range of item difficulties.

Finally, the study included students from both Singapore and Turkey. Although the combined dataset provided a large sample for calibration, possible cultural or educational differences between the two groups were not examined and warrants further investigation.

Future studies should encourage more challenge attempts per student, including a broader range of learner abilities, and examine subgroup differences to strengthen the measurement properties of the game system.

5 Conclusion

This study developed a Rasch-based game-leveling measurement scale for adaptive English vocabulary learning. Five vocabulary games were designed to reinforce high-frequency English words and calibrated using the Rasch Partial Credit Model. Gameplay data collected from students in Singapore and Turkey provided evidence for evaluating the measurement properties of the proposed game-leveling system.

The Rasch analysis established a stable hierarchy of game challenge difficulty, with high item separation and reliability indicating that the challenges were successfully calibrated across a broad range of difficulty levels. In contrast, the relatively low person separation and reliability suggest that the current gameplay data were less effective in distinguishing between learners of different ability levels, largely because of limited responses per student and the flexible nature of gameplay. Dimensionality analysis further indicated that the games predominantly measured a common underlying construct of English vocabulary ability.

Overall, the findings demonstrate the potential of Rasch analysis for developing evidence-based game-leveling systems for vocabulary learning. The results provide practical guidance for refining game content, improving learner targeting, and supporting adaptive educational game design. Future research may further refine the game-leveling measurement scale by improving the targeting of game challenges to learner ability levels and investigating whether the calibrated challenges function similarly across different groups of learners.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this paper.

References

1. Moody, S., Hu, X., Kuo, L., Jouhar, M., Xu, Z., Lee, S.: Vocabulary instruction: A critical analysis of theories, research, and practice. Educ. Sci. 8, 180 (2018). https://doi.org/10.3390/educsci8040180
2. Sedita, J.: Effective vocabulary instruction. Insights Learn. Disabil. 2(1), 33--45 (2005). Available at: https://keystoliteracy.com/wp-content/uploads/2012/08/effective-vocabulary-instruction.pdf
3. Nagy, W.E., Herman, P.A.: Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In: McKeown, M.G., Curtis, M.E. (eds.) The Nature of Vocabulary Acquisition, pp. 19--35. Erlbaum (1987)
4. Nation, I.S.P., Wang, M.K.: Graded readers and vocabulary. Read. Foreign Lang. 12, 355--379 (1999)
5. Ellis, R.: Principles of instructed language learning. Asian EFL J. 7(3), 9--24 (2005). Available at: http://www.asian-efl-journal.com/September_05_re.php
6. Dalton, B., Proctor, C.P., Uccelli, P., Mo, E., Snow, C.E.: Designing for diversity. J. Lit. Res. (2011). https://doi.org/10.1177/1086296X10397872

7. Mayer, R.E.: Computer games in education. Annu. Rev. Psychol. 70, 531--549 (2019). https://doi.org/10.1146/annurev-psych-010418-102744
8. Deci, E.L., Ryan, R.M.: The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. Psychol. Inq. 11(4), 227--268 (2000). https://doi.org/10.1207/S15327965PLI1104_01
9. Klimova, B.F.: Games in the teaching of English. Procedia Soc. Behav. Sci. 191, 1157--1160 (2015). https://doi.org/10.1016/j.sbspro.2015.04.312
10. Uberman, A.: The use of games for vocabulary presentation and revision. Forum 36(1) (1998). Available at: https://www.academia.edu/1082852/The_use_of_games_for_vocabulary_presentation_and_revision
11. Prensky, M.: Digital Game-Based Learning. Available at: http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Game-Based%20Learning-Ch5.pdf (2001)
12. Marzano, R.J.: Using games to enhance student achievement. Educ. Leadersh. 67(5), 71--72 (2010). Available at: http://www.ascd.org/publications/educational-leadership/feb10/vol67/num05/Using-Games-to-Enhance-Student-Achievement.aspx
13. Wouters, P., van Nimwegen, C., van Oostendorp, H., van der Spek, E.D.: A meta-analysis of the cognitive and motivational effects of serious games. J. Educ. Psychol. 105(2), 249--265 (2013). https://doi.org/10.1037/a0031311
14. Linacre, J.M.: A User's Guide to Winsteps/Ministep Rasch-Model Computer Programs, Version 5.9.2. Winsteps, Chicago (2025). Available at: https://www.winsteps.com/a/Winsteps-Manual.pdf
15. Linacre, J.M.: Optimizing rating scale category effectiveness. J. Appl. Meas. 3(1), 85--106 (2002)
16. Linacre, J.M.: Sample size and item calibration stability. Rasch Meas. Trans. 8(2), 370 (1994). Available at: https://www.rasch.org/rmt/rmt82.htm
17. Linacre, J.M.: What do infit and outfit, mean-square and standardized mean? Rasch Meas. Trans. 16(2), 878--879 (2002). Available at: https://www.rasch.org/rmt/rmt162f.htm
18. Bond, T.G., Fox, C.M.: Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 3rd edn. Routledge, New York (2015)
19. Linacre, J.M.: Reliability and separation of measures. Winsteps. Available at: https://www.winsteps.com/winman/reliability.htm
20. Mayer, R.E.: Computer Games for Learning: An Evidence-Based Approach. MIT Press, Cambridge (2014)
21. Linacre, J.M.: Dimensionality and structural validity investigation. Winsteps. Available at: https://www.winsteps.com/winman/multidimensionality.htm
22. Linacre, J.M.: Fit diagnosis: Infit, outfit, mean-square, standardized. Winsteps. Available at: https://www.winsteps.com/winman/misfitdiagnosis.htm

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

