The JALT CALL Journal
ISSN 1832-4215, 22, 103966 (2026)
https://doi.org/10.29140/jaltcall.2026.103966

Castledown

Differential effectiveness of text-to-speech technology on secondary school students' L2 speaking performance: A mixed-methods study

Hangyan Lu
Gratia Christian College, Hong Kong
luhangyan@gmail.com

Hong Sheung Chui
Gratia Christian College, Hong Kong
hschui@gratia.edu.hk

This study explored the impact of text-to-speech (TTS) technology on the English-speaking skills of secondary school students across different proficiency levels and investigated the factors influencing its effectiveness. The research employed a mixed-methods approach, combining pre- and post-test speaking assessments with semi-structured interviews and focus groups. Quantitative results revealed a pattern of differential effectiveness: the quasi-experimental group demonstrated significant gains overall when compared to their baseline, and between-group comparisons showed that this improvement was significantly greater than that of the control group across all four cumulative ability groupings. The effect was most pronounced among low-achieving students, while mid- and high-achieving students showed more modest gains. Qualitative findings elucidate this differential pattern, indicating that low-achieving students, who faced significant comprehension barriers and scarce learning resources, perceived TTS as a critical, compensatory tool. In contrast, high-achieving students found TTS misaligned with their goals for authentic communication and nuanced feedback. Although all students expressed a preference for human teachers, perceived usefulness was the primary driver of engagement. The findings suggest that, within the context of prepared read-aloud tasks, TTS technology may offer targeted support that could contribute to more equitable learning outcomes for low-achieving learners, though its effectiveness is mediated by learner proficiency, perceived task difficulty, and the availability of alternative resources. The results provide initial evidence that successful implementation

likely requires pedagogically differentiated strategies that align the tool's functionality with specific learner needs.

Keywords: text-to-speech (TTS), L2 speaking, students' perceptions, mixed-methods research, secondary education

Introduction

The integration of artificial intelligence (AI) in English language education has garnered increasing attention, with growing experimentation and research exploring its potential to transform learning experiences and enhance outcomes. Promoting inclusivity, or democratizing English language education (Tafazoli, 2024), is believed to be one of its major transformative potentials. However, much of this research has been concentrated in higher education settings (Crompton, 2023; Sharadgah & Sa'di, 2022; Yang & Kyun, 2022), leaving a gap in understanding its applicability and effectiveness in secondary schools. One reason for this gap may be related to the exam-driven nature of secondary education which is characterized by a rigid curriculum that leaves little room for experimentation. As new approaches usually involve a learning curve and carry uncertainty in outcomes, schools and teachers may linger on traditional approaches given the stakes of standardized assessment. Take Hong Kong as an example, despite policy shifts from traditional approaches such as grammar translation and oral-structural approach toward communicative and task-based learning approaches, traditional teaching methods such as dictations, grammar exercises and drills have remained as predominant practices in classrooms (Lau, 2020). This discrepancy between curriculum guidelines and classroom practices underscores the need for innovative, in-depth explorations of how AI can be effectively integrated into secondary school English education. As public examinations of English in Hong Kong primarily assess reading, grammar and composition, this study focuses on speaking skills as a pilot area to minimize washback effects and investigate AI's potential to enhance oral proficiency.

The success of AI integration depends on multiple factors, such as school readiness, teacher beliefs (Kohnke et al., 2024) and AI competencies, and students' AI literacies and ethics (NG et al., 2025). As early as 2023, the Education Bureau of Hong Kong (2025) launched an official AI curriculum that promotes responsible use of AI in upper primary and junior secondary school classrooms. While conversational AI tools like Google Assistant have shown promise in boosting learner's willingness to communicate (Waluyo & Pratiwi, 2025) and improving fluency and pronunciation (Eragamreddy, 2025), their effectiveness with young learners can be limited by challenges in keeping up with complex conversations. For secondary school students, a more structured and less cognitively demanding tool may be more practical.

AI-powered text-to-speech (TTS) tools may offer a simpler, more controllable alternative by providing clear, model-based input without the pressure of real-time conversation, making them suitable for fundamental speaking practice.

Accordingly, this study takes a focused step by examining the use of one TTS tool, Natural Reader, in a junior secondary speaking class. It aims to evaluate the tool's effectiveness in improving students' speaking performance while also exploring the reasons behind the outcomes by capturing stakeholder perspectives. Two research questions guide the investigation: (1) Can Natural Reader promote secondary school students' English-speaking performance? (2) What factors contribute to its effects – or lack thereof? By addressing these questions within Hong Kong's secondary education system, the research seeks to contribute empirical evidence on AI's role in transforming English language education in comparable contexts.

Literature review

Understanding ESL learners' English-speaking difficulties

Developing speaking proficiency in English as a second or foreign language (ESL/EFL) presents significant challenges for learners from diverse background. First, there are linguistic challenges. The most frequently reported difficulties are related to grammar, limited vocabulary, and pronunciation issues. These challenges are consistently identified among ESL/EFL learners across educational levels and in diverse contexts, including China, Indonesia, Malaysia, Oman, UAE, and the Philippines (Gan, 2013; Honsi, 2014; Jusnita, et al., 2024; Kulsum, et al., 2025; Labad et al., 2024; Normawati, et al., 2023; Suliman, 2020, Wardhani, et al., 2023; Yahaya et al., 2021). Psychological factors like anxiety, fear of making mistakes, lack of self-confidence, and lack of self-esteem are another type of common obstacles that discourage students from speaking and developing speaking competencies (Gan, 2013; Honsi, 2014; Kulsum, et al., 2025; Jusnita, et al, 2024; Separa, et al., 2020; Shen & Chue, 2019; Suliman, 2020; Yahaya et al., 2021; Zhang, 2009). Finally, traditional teaching methods and limited opportunities to use English in real-life contexts further exacerbate L1 interference and hinder L2 progress (Gan, 2013; Hosni, 2014; Sulima, 2020; Yahaya et al., 2021).

Interestingly, secondary school learners exhibit unique dynamics. Despite self-reported moderate speaking difficulties, many secondary school students can still perform well academically in English (Labad, et al., 2024), suggesting a disconnect between their communicative competence and exam performance. Additionally, a small-scale study in Indonesia found that secondary school students predominantly prefer kinesthetic learning style (85%), followed by auditory learning style (69%) and visual learning style (54%) (Jusnita & Ansar, 2024), raising questions about the alignment of teaching methods with learner preferences. While the sample sizes were limited in the two studies, these findings underscore the need for pedagogical and assessment innovations tailored to adolescents' cognitive and affective needs.

It is thus not surprising to find that learners expressed desire for varied and engaging speaking activities (Yahaya, et al., 2021). Technology-assisted solutions might offer a promising avenue to address these challenges. The following

section reviews the role of text-to-speech technology (TTS) in fostering learners' speaking skills, with a focus on their potential to mitigate linguistic, psychological, and pedagogical barriers.

TTS for developing speaking

Text-to-speech (TTS) technology, which converts written text into synthetic speech, has gained traction as a tool for developing English speaking skills, especially in contexts with limited exposure to English. Recent studies have explored its integration with speech recognition and mobile platforms to enhance pronunciation, fluency, and learner confidence (Moon, 2016; Moxon, 2024). While TTS has shown promise in various educational settings – from secondary schools (Promono, 2022) to universities (Meihami, 2013) – its effectiveness depends on addressing linguistic, psychological, and pedagogical barriers.

TTS can mitigate key linguistic challenges, such as pronunciation and fluency, by providing learners with repetitive, individualized practice. Regarding pronunciation, Nguyen and Vo (2024) found that shadowing TTS-generated speech improves students' word clarity, intonation, and memorization of sounds. For fluency in read-aloud tasks, which is also dependent on the speaker's familiarity with the vocabulary and reading comprehension, TTS has demonstrated measurable benefits. Chiang (2019) reported that students performed significantly better in TTS-led dictations than in teacher-led ones, which means they learn words better with the integration of TTS in preparing for and taking dictations, despite survey results in the same study suggesting that they did not appreciate TTS in the classroom. Similarly, Stodden et al. (2012) found that students who used TTS for an average of 40 minutes or more per week showed significant improvement in reading rate, vocabulary, and reading comprehension. However, a prevailing assumption in the literature is that TTS is particularly beneficial for struggling readers (e.g., Bone & Bouck, 2017; Stodden, et al, 2012), with recent work also focusing on beginner-level L2 learners (Saadia, 2023); yet this claim lacks robust comparative evidence involving mid- and high-achieving learners.

In addition, TTS reduces speaking anxiety by offering a low-pressure environment for practice, particularly for learners who fear judgment in classroom settings. Its accessibility supports self-directed learning (Van Lieshout & Cardoso, 2022) and foster greater confidence and interest in speaking English (Amin, 2024; Nguyen & Vo, 2024). Yet some learners still prefer human teachers for pronunciation modeling as synthetic voice may lack emotions (Chiang, 2019).

In exam-driven secondary education systems – where curricula often prioritize grammar and vocabulary drills over communicative practices – TTS offers a pragmatic solution to bridge the gap between exam performance and actual speaking competence. Research suggests such technology could particularly address L1 interference issues stemming from limited practice opportunities and vocabulary struggles (Krish & Oh, 2020), as it enables teachers to create tailor-made listening materials for various activities in class such as

dictation exercise (Oumaina, et al., 2018), interactive English listening (Sari, 2018), and more real-life communicative activities in class (Cardoso, 2022). It can also serve as a tool for self-directed L2 learning (Van Lieshout & Cardoso, 2022). However, realizing these affordances fully requires more contextualized research into implementation challenges, particularly regarding teachers' experiences with AI tools and their integration into classroom practices (Avrianti et al., 2026).

Methodology

Context and participants

The study was conducted among 63 Form 1 students (aged 12–13) from a public secondary school in Hong Kong, all enrolled in a mandatory "English speaking" course that offers a total of 35 minutes of class time per week over three months. The students, whose parents all provided informed consent, were organized into three parallel classes. All the three classes were taught by the same teacher, who is a native English speaker from Canada.

Research design

This is an explanatory sequential mixed-method study (Creswell & Creswell, 2018) with the first phase involving a quasi-experimental study to quantitatively examine the treatment effects and the second phase involving follow-up interviews to help explain in more detail the found treatment effects (or null effects). Quantitative research allowed for statistical analysis and objective results to answer RQ1 and to gather initial information for RQ2. Qualitative data collection methods were further employed to expand on the quantitative data previously collected on RQ2 and for a more comprehensive picture of the topic under study.

Instruments

In the first phase of TTS-based intervention, English speaking performance is operationalized as oral reading fluency and pronunciation accuracy, as measured by two read-aloud tasks. Practically, this operationalization is coherent with the developmental stage and curricular expectations of the Form 1 participants. For junior secondary L2 learners who are still developing foundational decoding skills, oral reading fluency shares substantial variance with broader speaking competence. The Education Bureau's (2018) junior secondary English curriculum guide explicitly advocates controlled speaking practice, such as shadow reading and read-aloud, as a means for students to improve their mastery of stress, rhythm and intonation and to transfer these prosodic patterns into their own spoken production. The read-aloud tasks thus assess the skills that the curriculum positions as foundational to speaking development at this educational stage. In addition, empirical evidence lends support to the use of

read-aloud tasks in assessing oral proficiency. Weir and Wu (2006) found that test forms containing read-aloud, answering questions, and picture description tasks were statistically parallel overall among intermediate-level EFL learners, indicating that read-aloud tasks can function comparably to more interactive task types in this context. Relatedly, Gráf et al. (2023) demonstrated a strong association between oral reading rate and CEFR-related speaking proficiency, further supporting the relevance of oral reading performance to broader speaking competence. Together, these findings suggest that read-aloud tasks, while not capturing all dimensions of speaking, can serve as a meaningful proxy for oral proficiency under certain conditions.

In the second phase, semi-structured interviews were conducted using one teacher interview protocol (see Appendix A) and one student interview protocol (see Appendix B). To ensure clarity during data collection, the term "English-speaking test" was used explicitly to refer to the two read-aloud tests implemented in the study, so that participants understood this phrase as denoting the specific assessment activity. However, when the interviews turned to more general discussion – for example, question 2's inquiry into students' preferences for learning speaking via Natural Reader or their teacher – the term "speaking" was conveyed to participants in a broader sense, encompassing all aspects of individual and interactive oral production. This dual usage allowed the qualitative data to capture perceptions of the read-aloud tasks specifically, while also exploring how teachers and learners perceived and used TTS across a wider range of English oral activities.

Phase 1: The quasi-experiment

The first phase of the study comprised of an experiment carried out between two classes of 42 students who were randomly chosen as the quasi-experimental group and the remaining class of 21 students who served as the control group. A same pre-test and post-test were administered to all participating students. Both tests involved asking students to read a piece of text aloud and the teacher graded students' reading performances based on a pre-set marking scheme. To prepare students for the pre-test, the teacher used traditional teaching methods with both groups (i.e., the test preparation period was directed by the teacher only). However, in preparing students for the post-test, the teacher continued to use traditional teaching method with the control group but introduced and used Natural Reader with the quasi-experimental group. Table 1 and Table 2 show the workflow.

Table 1. Study workflow with the control group

Pre-test stage (2 weeks)	Intervention stage (4 weeks)	Post-test stage (2 weeks)
1st week: 1. Teacher introduced pre-test excerpt with explicit instruction on key features 2. Students performed choral reading after teacher 3. Selected students read the excerpt individually to class	Weekly for a period of four weeks: 1. Teacher introduced a new excerpt with explicit instruction on key features 2. Students performed choral reading after teacher 3. Selected students read the excerpt individually to class	1st week: 1. Teacher introduced post-test excerpt with explicit instruction on key features 2. Students performed choral reading after teacher 3. Selected students read the excerpt individually to class
2nd week: 1. All students read the excerpt independently to class 2. Teacher graded each student's performance using a standardized marking rubric		2nd week: 1. All students read the excerpt independently to class 2. Teacher graded each student's performance using the same marking rubric as pre-test

Table 2. Study workflow with the quasi-experimental group

Pre-test stage (2 weeks)	**Intervention stage (4 weeks)**	**Post-test stage (2 weeks)**
1st week: 4. Teacher introduced pre-test excerpt with explicit instruction on key features 5. Students performed choral reading after teacher 6. Selected students read the excerpt individually to class	Weekly for a period of four weeks: 4. Teacher introduced a new excerpt with explicit instruction on key features 5. Teacher guided students' choral reading after Natural Reader 6. Teacher instructed and encouraged students to practice reading the excerpt individually with Natural Reader both in class and at home	1st week: 4. Teacher introduced post-test excerpt with explicit instruction on key features 5. Teacher guided students' choral reading after Natural Reader 6. Students practiced reading the excerpt individually with Natural Reader
2nd week: 3. All students read the excerpt independently to class 4. Teacher graded each student's performance using a standardized marking rubric		2nd week: 3. All students read the excerpt independently to class 4. Teacher graded each student's performance using the same marking rubric as pre-test

As Table 1 and 2 show, to prepare students for the pre-test, the teacher used traditional teaching method. He would put the pre-test excerpt on the board, give students explicit instructions such as drawing some intonation lines or underlining some key words that need to be stressed, and then invite the whole class to read after him, and then call out some of the students to read it individually to class. After the pre-test was completed, the teacher introduced Natural Reader to students and let it replace his role in leading students' choral reading in class and take up an additional role to assist students in their individual practice both in and outside class. Students were instructed and encouraged to experiment individually with the app's functions, using different accents and speeds for example.

Both pre-test and post-test excerpts were selected from the textbook that had been covered in other English lessons during the same semester. Consequently, the two excerpts were of similar difficulty and featured content familiar to students. Specifically, the two excerpts were chosen to be largely accessible, yet include a limited number of challenging words. This was intended to ensure baseline accessibility while incorporating elements that could capture a range of linguistic competencies. After each test, the teacher used a same marking rubric (see Table 3) to grade each student's performance. The rubric was also the one used for the passage reading component in the end-of-year speaking exam, or English paper 4 for Form 1 students.

Table 3. Marking rubric for pre-test and post-test

Description	Mark
No error and mispronounced words with good fluency	5
Some minor errors with moderate fluency	4
Able to follow and understand with errors and mispronounced words	3
Read in broken words and difficult to follow or understand	2
An attempt is made at some words. Cannot be understood	1
Nothing is read / No attempt	0

Quantitative data analysis procedures

Pre- and post-test data were obtained from teacher assessments and analyzed using SPSS Version 31.0. To measure the effect of Natural Reader in enhancing speaking performance (RQ1), descriptive statistics and an independent-samples t-test on pre-test scores were first conducted to establish baseline comparability between the quasi-experimental and control groups. Subsequently, descriptive statistics and one-sample t-tests were used to determine whether each group's mean value-added score (post-test score minus pre-test score) differ significantly from zero, indicating whether significant gains had occurred after the intervention. In addition, an independent-samples t-test comparing valued-added mean scores between the two groups for the full sample was conducted to assess the overall between-group effect. As this overall comparison did not reach statistical significance ($p = .096$), which may partly reflect a ceiling effect of students who scored the maximum 5 at pre-test, further exploratory subgroup analyses were undertaken. Specifically, independent-samples t-tests comparing value-added scores between the quasi-experimental and control groups were conducted across four cumulative ability groupings: students with pre-test scores of 0–1, 0–2, 0–3, and 0–4.

Phase 2: Interviews

Phase 2 employed semi-structured interviews to elaborate on the quantitative findings from Phase 1 and address RQ2. Using a stratified sampling approach,

students from the quasi-experimental group were evenly divided into low-, mid-, and high-achieving subgroups based on their pre-test scores. Data collection consisted of a 30-minute individual interview with the teacher and three 15-minute focus groups with students from the quasi-experimental group (one each for low-, mid-, and high-achieving students; n=4 per group). This design was implemented specifically to investigate the differential effectiveness revealed by the Phase 1 results, wherein low-achieving students demonstrated the most pronounced improvement, while mid- and high-achieving students also showed gains, though more modest in magnitude.

To ensure clarity and allow for reflective preparation, interview protocols were shared with the teacher and the selected student participants approximately one week in advance. The teacher interview was conducted in English. The student focus groups were conducted flexibly in English, Cantonese and Mandarin, enabling student participants to fully comprehend the questions and articulate their perspectives. All sessions were audio-recorded with prior consent from the teacher and from the parents of the student participants. The recordings were subsequently transcribed verbatim to ensure data accuracy.

Qualitative data analysis procedures

The qualitative data was organized and managed systematically using NVivo software (Version 11 Plus). Following Braun and Clarke's (2006) thematic analysis framework, all interview and focus group transcripts were imported into the software for analysis. The process began with an initial phase of familiarization and manual open coding, whereby key phrases and recurring ideas were highlighted and assigned preliminary codes. These codes were then reviewed and grouped into potential themes based on conceptual similarities. The themes were further refined and defined through an iterative process, examining relationships within and across the data set.

The final themes were interpreted in direct relation to the research questions, focusing on similarities and differences in perspectives across the three student groups (high-, mid-, and low-achieving) and between the students and the teacher. To ensure findings remained grounded in the data, representative quotations were selected to illustrate key findings. This structured analytical approach facilitated a nuanced understanding of how Natural Reader was perceived and appropriated by different users and its perceived impact on speaking performance (RQ2).

Findings

Phase 1 quantitative findings

After cleaning incomplete data, the results of 57 students were analyzed. There were 38 (66.7%) boys and 19 (33.3%) girls in total, 27 boys and 11 girls from the quasi-experimental group, and 11 boys and 8 girls from the control group.

Prior to examining the effects of the intervention, independent-samples

t-tests were conducted on pre-test scores to assess whether the quasi-experimental group and the control group differed at baseline. Table 4 shows the results which indicated no statistically significant difference between the two groups, t(55) = −.736 and p value at .465, with a negligible effect size (Cohen's d = −.207). Descriptive statistics further show comparable distributions: skewness values were −.171 for the quasi-experimental group and .173 for the control group, while Kurtosis values were −1.161 and −1.047, respectively. These findings suggest that the two groups were reasonably comparable in their English-speaking performance prior to the intervention.

Table 4. Descriptive statistics and independent-samples t-test for pre-test scores

Pre-test score	Quasi-experimental group	Control group
Mean score	2.89	2.58
Standard deviation	1.641	1.261
Standard error	.266	.289
Kurtosis	−1.161	−1.047
Skewness	−.171	.173
Total number	38	19
Maximum score	5	5
Minimum score	0	1
Independent-samples t-test: t-value (p value; Cohen's d estimate)	−.736; p = .465 d = −.207	

Table 5 presents the descriptive and inferential statistics for value-added scores (post-test minus pre-test) for both groups. Both groups demonstrated positive mean value-added scores, indicating that some improvement occurred in each condition. However, the magnitude and statistical significance of these gains differed substantially between groups. The quasi-experimental group (n = 38) achieved a main gain of .4211 (SD = 0.975), which was statistically significant from zero, t(37) = 2.659, p = .012, with a medium effect size (Cohen's d = .431). In contrast, the control group (n = 19) showed a negligible mean gain of .0526 (SD = 1.025), which did not differ significantly from zero, t(18) = .224, p = .826, d = .05. The distribution of gain scores in the quasi-experimental group exhibited positive skewness (.510) and kurtosis (.916), suggesting that while most participants improved, a subset showed more pronounced gains. In the control group, skewness (−.115) and kurtosis (−.554) indicated a more symmetrical, flatter distribution of change scores. Taken together, these results indicate that while both groups experienced improvement over the study period, only the quasi-experimental group demonstrated statistically and practically significant gains, supporting the effectiveness of the TTS-based intervention.

Table 5. Descriptive statistics and t-test results for valued-added scores

Pre-test score	Quasi-experimental group	Control group
Mean score	.4211	.0526
Standard deviation	.975	1.025
Standard error	.158	.235
Kurtosis	.916	−.554
Skewness	.510	−.115
Total number	38	19
Maximum score	3	2
Minimum score	−2	−2
One-sample t-test (vs. zero): t-value (p value; Cohen's d estimate)	2.659 (p = .012; d = .431)	.224 (p = .826; d = .051)
Independent-samples t-test (between groups): t-value (p value; Cohen's d estimate)	−1.321 (p = .096, d = −.371)	

Notably, when the two groups are compared together directly using an independent-samples t-test on mean valued-added scores, the between-group difference did not reach statistical significance (p = .096, d = −.371) (see Table 5). This null overall result, however, may partly reflect the inclusion of students who scored the maximum of 5 at pre-test and therefore had no room for measurable improvement. To examine whether the intervention effect varied by learners' initial proficiency once these ceiling-affected cases were accounted for, we conducted a series of independent-samples t-tests on value-added scores between the quasi-experimental and control groups across four progressively inclusive ability groupings (see Table 6).

Table 6. Independent-samples t-tests for value-added scores between quasi-experimental and control groups by pre-test ability grouping

Ability grouping	Pre-test scores included	Quasi-experimental group mean gain	Control group mean gain	P value (*significance at 0.05)	Cohen's d
1	0–1	1.111	0.0	.007*	−1.291
2	0–2	.6875	−.111	.032*	−.811
3	0–3	.7826	−.0714	.010*	−.832
4	0–4	.5862	.0556	.050*	−.504

Note. As these subgroup analyses involve overlapping samples and multiple comparisons, they should be interpreted as exploratory and pattern-seeking. The consistent direction and significance of effects across groupings, rather than any single comparison, is the primary finding of interest.

Three clear patterns emerged. First, the quasi-experimental group consistently outperformed the control group across all ability groupings, with statistically significant differences observed at every level (all p ≤ .05). Second, the largest between-group difference was observed in the lowest ability grouping (pre-test

scores of 0–1), where the quasi-experimental group achieved a mean gain of 1.11 compared to 0.00 in the control group ($p = .007$, $d = -1.291$). This represents a large effect and suggests that the intervention was particularly impactful for the least proficient learners. Third, as higher-achieving students were progressively included, the quasi-experimental group's mean gain dropped and then stabilized at a moderate level: .6875 in Grouping 2 (scores 0–2, $p = .032$, $d = -.811$), .7826 in Grouping 3 (scores 0–3, $p = .010$, $d = -.832$), and .5862 in Grouping 4 (scores 0–4, $p = .050$, $d = -.504$). All between-group differences remained statistically significant. This pattern suggests a threshold effect: the intervention effect conferred a pronounced advantage on the low-achieving students, while its benefit for students above this lowest level was more modest but consistently present across the remaining proficiency bands. On the other hand, the consistently negligible gains in the control group across all ability levels – remaining close to zero and turning slightly negative in Grouping 2 (−.111) and 3 (−.0714) – further confirm that maturation or extraneous exposure alone cannot account for the quasi-experimental group's improvement.

Phase 2 qualitative findings

The qualitative findings from Phase 2 both triangulate the quantitative results for RQ1 – concerning the effectiveness of Natural Reader in promoting secondary school students' English-speaking performance – and to address the explanatory aims of RQ2, which examines the factors underlying the observed treatment effects (or null effects). In the following section, the qualitative findings will be presented from the perspectives of the teacher, then the high-, mid-, and low-achieving students. All names cited are pseudonyms.

Teacher's perspective

The teacher's perspective triangulated and elaborated on the qualitative findings. He perceived Natural Reader as a valuable pedagogical complement but emphasized its role as a supplement rather than a replacement for direct instruction. He validated the overall improvement through its utility for pronunciation practice ("useful for understanding stress and intonation") and accessibility ("portable, they can [use Natural Reader] at school, home, wherever"). However, he concurrently noted a critical limitation in student engagement ("I think right now, they [students] were a little bit more engaged when it was the teacher, the traditional method").

A central insight from the teacher directly explained a key quantitative result: the differential gains by ability group. He identified the constrained 0–5 grading scale as creating a methodological ceiling for higher-proficiency students: "It is harder to go from a four to a five than from a one to a two … there is a ceiling there." This explicit recognition of a ceiling effect aligns with the statistically insignificant gains observed for low-achieving students and the non-significant gains for their mid- and high-achieving peers.

The interview further revealed contextual and practical factors that could

contribute to mitigated effects. Technical implementation hurdles, including initial "technical difficulties" and a significant learning curve, were noted. The teacher estimated a normalization period of approximately four classes, or about one month, before the tool could be used fluently. He also speculated that sustained use would likely be limited to self-motivated students and questioned the efficacy of self-correction at this proficiency level, commenting, "I think at this stage the teacher correction would be a bit more beneficial than the self-correction just because I think a lot of students wouldn't be sure what to correct or when to correct it by themselves yet."

Finally, systemic barriers within the exam-driven curriculum were highlighted, with lessons described as "packed with exam-related stuff." To enhance future implementation, the teacher proposed strategic parent involvement to encourage home use and advocated a hybrid pedagogical model that balances TTS tool use with interactive, teacher-led tasks.

High-achieving students' perspectives

The perspectives of high-achieving students provide a nuanced explanation for the more modest gains observed within this subgroup. Their feedback suggests both the source of the improvement and the factors that constrained its magnitude.

The foundation for any gain lies in their engagement with the tool within the teacher's structured classroom context. All four students acknowledged the conceptual value of hybrid instruction, with one summarizing that "teacher plus Natural Reader is better". This sanctioned in-class use under pedagogical guidance represents the primary mechanism through which the intervention could influence their performance.

However, three interconnected factors from the student perspectives effectively capped this gain, preventing a more significant quantitative effect. First, a consistent preference for human interaction limited their engagement with the tool. All four students consistently emphasized the teacher's superiority in delivering natural pronunciation, emotional expressiveness, and real-time feedback. As one student noted, "[My teacher] is better than AI because he can teach pronunciation more accurately," while another added that the teacher could convey emotions, making the experience more fun and like a real person. This preference for human interaction was rooted in a desire for genuine communication: "You have face-to-face [interaction], like when you go to another country, somebody just [talks] face-to-face with you."

Second, there was a complete absence of voluntary practice outside of class. This behavioral disengagement was driven by their perception that the core speaking tasks were not challenging ("the test is very easy") and their confidence in their existing proficiency ("my English is very good"). As one student stated, "I don't use Natural Reader except in the classroom," a sentiment unanimously echoed within the group. Their self-regulated practice relied instead on preferred alternatives, such as conversing with fluent English-speaking peers or family members.

Third, even within the classroom setting, half of the students critically questioned the tool's pedagogical utility for their level. While conceding the abstract hybrid model, two ultimately rejected it after deliberation ("I suggest maybe just teacher"; "I prefer only teacher"). Their critique centered on TTS's inability to provide contextual and corrective guidance ("the teacher can tell you where to stop, where to continue. But if you use Natural Reader, you don't know where to stop") necessary for refining advanced skills.

In summary, this student perspective elucidates the observed outcomes: the in-class teacher-facilitated use of Natural Reader provided a baseline for the positive trend, while the convergence of pedagogical preference, external disengagement, and perceived tool limitations explains why this trend remained modest. Their feedback further implies that for TTS to demonstrate significant efficacy with high-achieving learners, it may need to be coupled with more advanced, interactive functionalities and more cognitively demanding tasks that align with their learning needs.

Mid-achieving students' perspectives

The perspectives of mid-achieving students provide a clear contrast to those of high-achieving students, helping to explain why both subgroups experienced only modest gains, and revealing both shared and distinct reasons underlying this pattern. Both groups shared a preference for the teacher in class (expressed by all four high-achieving students and all four mid-achieving students; e.g., from this group: "I just feel more used to teacher's reading"] and a general disengagement in using Natural Reader outside of class (reported by all four high-achieving students and three of the four mid-achieving students); however, the underlying dynamics of engagement differed.

A key differentiating factor was the nature of their resistance. Unlike high-achieving students, mid-achieving students did not criticize the tool for failing to meet advanced pedagogical needs. Instead, their primary barriers were technical friction and a lack of habitual English engagement. Two of the four students in this group cited technical friction, describing them as "a bit complicated" and "not easy to use." One acknowledged a lack of habitual English engagement: "I rarely read outside class". This suggests that their hurdles were usability and practice habit, rather than a ceiling in task difficulty or a critique of AI feedback limitations. Consequently, the structured, in-class use mandated by the study design provided a consistent dose of accessible pronunciation practice – described as "pretty difficult" due to unfamiliar vocabulary by one student "okay" by two others – that was largely beneficial but lacking in their existing routines.

Also notably, one student in this group reported some positive experimentation, acknowledging the tool's accurate pronunciation and accent customization. This indicated that when technical barrier was momentarily overcome, the tool's core utility for modeling was perceptible. However, this did not translate into committed usage, as it was outweighed by the stronger forces of technical friction and established reliance on familiar human support systems.

Each of the four students in this group identified a distinct source of support: friends ("ask my friend to teach me how to read"), family members ("I will ask my parents; I will practice with my parents"), a domestic helper ("I got an auntie to take care of my brother and me. So every day I need to use English, just speak to her"), and a private tutor (one student reported twice-weekly sessions with a private English tutor).

In summary, the greater gains among mid-achieving students stem from the tool's alignment with an unmet need. For them, mandatory in-class use provided consistent, quality pronunciation practice lacking in their routines.

Low-achieving students' perspectives

The perspectives of low-achieving students provide a definitive explanation for the pronounced quantitative gains observed in this subgroup, completing the explanatory model for the differential effectiveness of the intervention across proficiency levels. While all subgroups preferred human teachers for emotional expressiveness ("teacher's reading is with more emotions"), low-achieving students' critiques revealed a fundamental barrier of comprehension. Three of the four students in this group concurred that they could understand only about one-third of teacher instruction (e.g., "I don't understand what the teacher says […] I can understand about 1/3"), in contrast to the technical or pedagogical resistance of their peers. For these students, when approximately two-thirds of instruction was inaccessible, a supplemental tool like TTS could offer little benefit without accompanying linguistic scaffolding.

This comprehension gap, compounded by the task's challenging nature for half the group (two students described the tests as "a little bit difficult" due to vocabulary gaps and pronunciation challenges, while the other two found them "okay"), was exacerbated by severe resource scarcity. Outside of class, three of the four students in this group mentioned classmates as their only human resource. Among these three, two found peer support to be unreliable: one noted that classmates' translation skills were not always dependable ("sometimes you ask the wrong person"), while the other pointed to limited availability ("sometimes the classmate who can help is on leave"). The fourth student in the group expressed a more general sense of helplessness ("I just listen. It is not working. I can't think of any good method to improve [my speaking]".

In this context, Natural Reader served as a unique, compensatory tool: it provided on-demand, comprehensible and repeatable pronunciation input that was otherwise absent. For one outlier student who engaged with it daily for two months with self-purchased materials ("read after Natural Reader … almost every day"), it became a transformative, facilitating significant vocabulary gains and test score improvement. The other three students, who did not understand teacher's instruction and encouragement of using Natural Reader in the first place, expressed openness to try TTS post-interview ("I may try it when I have time"; "I would like to try it").

Therefore, the pronounced gains in this subgroup can be attributed to a

powerful alignment: the tool's core functionality directly addressed their most acute and unmet needs – comprehensive modeling and reliable practice access.

Discussion

This study provides a nuanced answer to whether Natural Reader can promote secondary school students' English-speaking performance on prepared read-aloud tasks. Quantitatively, the intervention in the quasi-experimental group yielded a positive value-added score higher than that in the control group, indicating an overall benefit in improving speaking performance within this sample of 57 secondary school students. This overall gain, despite the relatively short two-month intervention, is noteworthy when considered alongside prior research on the timeline for TTS benefits to emerge. Suro and Ono (2016) found that a brief, three-week TTS intervention was insufficient for benefits to materialize, with their control group outperforming their experimental group. The teacher in the present study similarly observed a necessary "learning curve", estimating that approximately four sessions (or one month) were needed for students to normalize using the new tool before its benefits could begin to surface. That our study nonetheless detected positive gains within two months suggests that the intervention may have been beginning to take effect within the window captured. Had the intervention been extended, it is plausible that the gains would have been larger, as the learning curve would have been more fully traversed. The present findings may therefore be interpreted as a conservative estimate of the intervention's potential.

More specifically, the quantitative findings provide preliminary evidence suggesting TTS's effectiveness as a targeted intervention for low-achieving students, who demonstrated significant gains in pronunciation and fluency. Our qualitative findings corroborate with prior research, validating that TTS supports this subgroup through repetitive, individualized practice (Nguyen & Vo, 2024) and improves reading rate, vocabulary, and reading comprehension (Stodden et al., 2012), underscoring its role in mitigating linguistic barriers where baseline gaps exist.

The current study also identifies two major factors influencing TTS effectiveness on developing passage-reading skills: prior competence and resource access. With low prior competence in reading a passage, low-achieving students in our sample faced significant baseline knowledge gaps creating greater marginal gains potential. In contrast, high-achieving peers tend to face ceiling effects from saturated skills. Resource access also plays a major role. For low-achieving students, their limited external support positions TTS as a vital scaffold. Conversely, high-/mid-achieving students leverage stable human networks, diminishing TTS's perceived utility. These factors help explain the pronounced improvement unique to this cohort – an outcome not evident to the same degree among higher-achieving peers with saturated skills or stable support networks.

On the psychological level, TTS adoption appears to depend not on technical perfection but on contextually salient benefits. The current study shows

that TTS's potential to engage students with its affordance of providing repetitive, individual practice anytime anywhere is differentially realized. Low-achieving students demonstrated higher motivations to engage with TTS after class; mid-achieving students showed low interest, and high-achieving students just rejected their interest. This differential motivation to use TTS after class reveals how psychological barriers to speaking practice, if any, are navigated through perceived utility. While all subgroups voiced the documented limitations in TTS naturalness and expressiveness (Moon, 2012; Chiang, 2019), this did not determine TTS adoption. Although mid-achieving students reported "technostress" (Kohnke et al., 2024) – citing technical complexity as a major barrier – high-achieving students did not express similar concern yet still show no interest in using TTS, providing further evidence that perceived learning gains stood as the critical motivator. Low-achieving students demonstrated higher motivation to use TTS outside class, with one student reported sustained daily use over two months and attributed significant vocabulary and test score gains to the tool. Others in this cohort expressed openness to adoption after witnessing these benefits. This aligns with Davis's (1989) technology acceptance model, wherein perceived usefulness significantly outweighs ease of use in driving usage behavior. Notably, within our sample, low-achieving students' resource scarcity and baseline gaps appear to have amplified their perception of TTS's usefulness, whereas mid-/high-achieving students – facing ceiling effects and possessing alternative supports – perceived more modest gains.

High-achieving students' disengagement with TTS in this study stemmed not demonstrably from tool inefficacy but rather from goal misalignment. Quantitative ceiling effects – evidenced by non-significant gains and student assertions that "the test is easy" – converged with qualitative insights: these learners prioritized authentic communicative competence over exam-focused pronunciation drills. As one high-achieving student emphasized, the ability to engage in genuine interaction (e.g., talking face-to-face in a foreign country) represented their core unmet need. This finding echoes Labad et al.'s (2024) discovery of examination-competence disconnect, wherein strong academic performance masks undeveloped speaking skills and latent anxiety among secondary school students. Notably, the exam-centric design of TTS implementation in this study optimized phonological accuracy and fluency rather than interactive adaptability.

Taken together, the results provide preliminary indications that differentiated support, real-time orchestration, and enhanced peer support may be beneficial for integrating AI tools such as TTS in language teaching. This interpretation responds to the divergent needs uncovered across proficiency levels within our sample. For low-achieving students, who struggle to comprehend teacher discourse, TTS must be coupled with scaffolding mechanisms that address fundamental linguistic barriers. This could include embedded translation support, slower speech rates, or synchronized visual text to reinforce auditory input. Such adaptations could augment the tool's value as a comprehensible input source, enabling these learners to access content that would otherwise remain out of reach.

For high-achieving students, while curricula constraints may limit systemic exam reform to mitigate ceiling effects found in the current study, TTS may still serve as a viable scaffold for communicative practice. Specifically, its capacity to generate varied accents and adjustable speech rates, combined with other tools such as Chatbots and speech recognition apps, could potentially enable teachers to design in-class simulations of real-world communication (i.e., role-plays with regionally authentic dialogues such as student conversations on campus, travel situations, or customer service interactions) that move beyond exam-focused drilling, thereby potentially bridging the competence-examination gap without compromising assessment frameworks. By incorporating physical response tasks, role-plays, and peer interactions, these activities might also align with secondary school students' predominant kinesthetic and auditory learning preferences (Jusnita & Ansar, 2024).

While students across different proficiency levels enjoy different resources, they all in our study demonstrated a natural reliance on peers like friends and classmates. This finding points to a potential opportunity to enhance TTS effectiveness through structured collaborative frameworks. Rather than relying on ad hoc and some fragile peer assistance (e.g., depending on classmate's availability or translation skills), instructors might consider deliberately designing mixed-proficiency group activities when integrating TTS as a collaborative tool. For example, teachers could assign small groups with diverse skills to co-create dialogue scripts, then practice role-plays using varied TTS accents, or jointly complete pronunciation challenges where the tool provides model outputs. Preliminary evidence from this study suggests that by doing this, low-achieving students may benefit from linguistic scaffolding and immediate feedback from peers, while mid- and high-achieving students might deepen their metacognitive and explanatory skills through teaching others. This approach has the potential to contribute to an inclusive learning environment, but these possibilities require further empirical validation across diverse contexts.

Conclusion and future research

The study investigated the effectiveness of TTS technology in improving the English-speaking skill of secondary school students across varying proficiency levels on prepared read-aloud tasks. The findings suggest a pattern of differential effectiveness: TTS appeared to serve as a potentially powerful compensatory tool for low-achieving students, facilitating pronounced gains in pronunciation and fluency by addressing their lower baseline proficiency and resource scarcity. Among mid- and high-achieving students, the intervention also yielded gains, though these were more modest in magnitude. This divergence underscores that technological affordances alone are insufficient; effectiveness is mediated by learners' prior competence, perceived usefulness, and access to alternative support. Additionally, the widespread reservations regarding TTS's artificiality across all groups, coupled with high-achieving students' expressed need for authentic communicative practice, challenges the notion of a one-size-fits-all AI solution. Consequently, these preliminary findings suggest

that the successful integration of TTS may depend on pedagogically informed, differentiated implementation – one that leverages its capacities for scaffolded support for low-achieving learners while repurposing its features for communicative competence development for high-achieving students. Ultimately, this study provides initial evidence that technology's role in language education is not to replace human-centric instruction but to complement it in ways that are contextually responsive and potentially address the diverse needs within a classroom.

Several limitations in the study design and implementation warrant acknowledgement while pointing towards productive avenues for future research. First, read-aloud tasks do not capture interactive, strategic and discourse-level aspects of L2 speaking. Second, the 0-5 grading scale may have lacked sufficient sensitivity at the upper end to capture improvement among high-achieving learners. Third, speaking performances were rated by a single rater, and inter-rater reliability was not established. Fourth, the duration and scope of the study may also have constrained the observable impacts of TTS integration. Future research could address these limitations by incorporating spontaneous or interactive speaking tasks, employing a more finely graduated assessment instrument, using multiple raters with established inter-rater reliability, and implementing longer interventions with larger, more diverse samples. Such designs could capture longitudinal effects on motivation, habitual use, and sustained linguistic gains across proficiency levels.

In addition, the current study underscores the critical role of pedagogical support and curricular alignment. Teacher training emerged as a potentially vital factor – not only for technical instruction but also for designing accessible, communicative activities that leverage TTS flexibility. Based on these preliminary findings, future implementations may benefit from providing educators with professional development focusing on differentiated TTS orchestration, such as creating tasks completed by mixed-ability groups or designing simulated real-world scenarios using accent and speed modulation. Additionally, research should explore how TTS can be embedded within existing exam-oriented systems to promote communicative competence without sacrificing academic performance.

Finally, while this study provides preliminary evidence of TTS's compensatory value for low-achieving students within this sample, the more modest gains among mid- and high-achieving students highlight a broader concern that warrants further investigation. Our findings resonate with Hao et al. (2025), who observed a similar pattern in an AI-agents-supported online course: learners with low prior knowledge demonstrated the greatest learning gains, while those with high prior knowledge experienced learning loss. That this pattern emerges across such different technological contexts – a simple TTS tool and a sophisticated AI-agents-supported online course – tentatively suggests that prior competence may mediate learning gains regardless of technological complexity. In our study, high-achieving students as a group did not regress, but two of the ten students at the pre-test ceiling did decline, as reported in the findings. This parallel underscores a potential risk: undifferentiated technology

integration may yield diminishing returns for or even harm high-achieving learners. These results complicate the narrative of AI's potential for "democratizing English language education" (Tafazoli, 2024), suggesting that without deliberate pedagogical scaffolding, AI tools may inadvertently create new educational inequalities. To truly realize Tafazoli's vision of "promot[ing] inclusive education by providing tailored experiences that accommodate diverse cultural backgrounds and learning needs" (p. 9), future implementation must address not only access but also differentiation, with the goal of ensuring that high-achieving students are adequately challenged while low-achieving learners are supported. Future research should explicitly investigate compensatory scaffolds for high-achieving students, particularly those susceptible to ceiling effects.

Use of generative AI

This manuscript made use of DeepSeek to assist in paraphrasing and improving the clarity of text. No AI tools were used for the generation of research hypotheses or the writing of conclusions. The final manuscript was authored and reviewed by the listed authors without additional AI input.

References

Amin, E. A.-R. (2024). EFL students' perception of using AI text-to-speech apps in learning pronunciation. *Migration Letter, 21*(3), 887–903. https://migrationletters.com/index.php/ml/article/view/6905/5256

Avrianti, N., Yuliana, Y. G. S., & Ikhsanudin. (2026). AI-based pronunciation training tools for English language learners: A systematic review. *Journal of English Education Program, 7*(1), 118–134. https://doi.org/10.26418/jeep.v7i1.99864

Bone, E. K., & Bouck, E. C. (2017). Accessible text-to-speech options for students who struggle with reading. *Preventing School Failure: Alternative Education for Children and Youth, 61*(1), 48–55. https://doi.org/10.1080/1045988X.2016.1188366

Chen, K. T. C. (2022). Speech-to-text recognition in University English as a foreign language learning. *Education and Information Technologies, 27*, 9857–9875. https://doi.org/10.1007/s10639-022-11016-5

Chiang, H. H. (2019). A comparison between teacher-led and online text-to-speech dictation for students' vocabulary performance. *English Language Teaching, 12*(3), 77–93.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Education Bureau, The Government of the Hong Kong Administrative Region. (2018). *Supplement to the English language education key learning area curriculum guide (Secondary 1-3).* https://www.edb.gov.hk/attachment/en/curriculum-development/kla/eng-edu/Curriculum%20Document/ELE%20JS%20Supplement%202018.pdf

Education Bureau, The Government of the Hong Kong Administrative Region. (2025, January 21). *Innovation and Technology Education.* https://www.edb.gov.hk/en/curriculum-development/kla/technology-edu/resources/InnovationAndTechnologyEducation/resources.html

Eragamreddy, N. (2025). Exploring the role of Google Assistant in enhancing EFL learners' speaking skills: A systemic review. *The JALT CALL Journal, 21*(3), 2223. https://doi.org/10.29140/jaltcall.v21n3.2223

Gan, Z. (2013). Understanding English speaking difficulties: An investigation of two Chinese populations. *Journal of Multilingual and Multicultural Development, 34*(3), 231–248. http://dx.doi.org/10.1080/01434632.2013.768622

Gráf, T., Huang, L., & Cilibrasi, L. (2023). Oral reading tasks as proficiency indicators: Insights from a learner corpus study. *International Journal of Learner Corpus Research, 9*(2), 155–179. https://doi.org/10.1075/ijlcr.22013.gra

Hao, Z., Cao, J., Li, R., Yu, J., Liu, Z., & Zhang, Y. (2025). Mapping student-AI interaction dynamics in multi-agent learning environments: Supporting personalised learning and reducing performance gaps. *Computers & Education, 241*, 105472. https://doi.org/10.1016/j.compedu.2025.105472

Honsi, S. A. (2014). Speaking difficulties encountered by young EFL learners. *International Journal on Studies in English Language and Literature, 2*(6), 22–30.

Jusnita, N., Arismunandar, & Ansar, H. (2024). The difficulties of speaking encountered by the students at English course. *Proceedings of International Conference on Education, 2*(1), 528–535. TIIKM Publishing

Kohnke, L., Zou, D., & Moorhouse, B. L. (2024). Technostress and English language teaching in the age of generative AI. *Educational Technology & Society, 27*(2), 306–320. https://doi.org/10.30190/ETS.202404_27(2).TP02

Krish, P. & Oh, C. M. (2020). A case study of L1 interference in speech acts among Chinese L2 students. *The Southeast Asian Journal of English Language Studies, 26*(1), 106–118. http://doi.org/10.17576/3L-2020-2601-08

Kulsum, E. M., JamJami, A. Y., Uyun, A. S., Widianingsih, A. W. S. (2025). A study on the difficulties in English speaking for EFL students. *Journal on Education, 7*(2), 10648–10659.

Kumar, Y., Singh, C., & Koul, A. (2022). A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. *Multimedia Tools and Applications, 82*, 15171–15197. https://doi.org/10.1007/s11042-022-13943-4

Labad, C. M. B., Rabena, R. P., Garduce, S. J. S., & Neri, E. (2024). Speaking difficulties and academic performance in English among junior high school students. *International Journal of Research and Innovation in Social Science, VIII*(IIIS), 2645–2656. https://dx.doi.org/10.47772/IJRISS.2024.803192S

Meihami, H. (2013). Text-to-speech software: A new perspective in learning and teaching word stress, word intonation, pitch contour, and fluency of English reading. *International Letters of Social and Humanistic Sciences, 8,* 24–33.

Moon, D. (2012). Web-based text-to-speech technologies in foreign language learning: opportunities and challenges. In T. Kim, J. Ma, W. Fang, Y. Zhang, & A. Cuzzocrea (eds.). *Computer applications for database, education, and ubiquitous computing,* (pp. 120–125). Springer.

Moon, D. (2016). Development and evaluation of an English speaking task using smartphone and text-to-speech. *The Journal of the Institute of Webcasting, Internet and Telecommunication, 16*(5), 13–20.

Moxon, S. (2024). All-talk: Enhancing EFL pronunciation with Microsoft Azure speech services. *ABAC Journal, 44*(4), 139–161. https://doi.org/10.59865/abacj.2024.58

Ng, D. T. K., Chan, E. K. C., & Lo, C. K. (2025). Opportunities, challenges and school strategies for integrating generative AI in education. *Computers and Education: Artificial Intelligence, 8,* 100373. https://doi.org/10.1016/j.caeai.2025.10037

Nguyen, T. P., & Vo, T. K. N. (2024). Impacts of text shadowing in developing EFL young learners' speaking skills. *Journal of Knowledge Learning and Science Technology, 3*(1), 284–299. https://doi.org/10.60087/jklst.vol3.n1.p299

Normawati, A., Nugrahaeni, D. A., Manggolo, N. S. K. H., & Susanto, A. I. F. (2023). EFL learners' difficulties in speaking English. *English Language and Education Spectrum, 1*(1), 1–9.

Oumaima, Z., Abdelouafi, M., & Meryem, E.H. (2018). Text-to-speech technology for Arabic language learners. In *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)* (pp. 432–436). IEEE. https://doi.org/10.1109/CIST.2018.8596372

Pramono, T. P. R. P. (2022). Improving students' ability in speaking of procedural text with Balabolka ('text to speech'). *ELite Journal: International Journal of Education, Language, and Literature, 2*(2), 73–78. https://doi.org/10.26740/elitejournal.v2n2.p73-78

Pratiwi, D. I. (2025). AI chatbot-assisted English learning and willingness to communicate: A narrative meta-synthesis of evidence from Asian English as a foreign language context. *The JALT CALL Journal, 21*(3), 102884. https://doi.org/10.29140/jaltcall.v21n3.102884

Rajendran, M., Subramanian, J., Ramesh, S., Lamessa, O., Prakash, A., & Revathi, P. (2024). Investigation into the source of students' English speaking difficulties. *World Journal of English Language, 14*(3), 453–462. https://doi.org/10.5430/wjel.v14n3p453

Saadia, K. H. (2023). *Assessing the effectiveness of text-to-speech and automatic speech recognition in improving EFL learner's pronunciation of regular past-ed.* [Unpublished master's thesis]. University of Biskra.

Separa, L. A. C., Generales, L. J., & Medina, R. J. S. (2020). Situational speaking difficulties of English as second language learners in the Philippines. *Journal of Southeast Asian Studies, 25*(1), 144–167. https://doi.org/10.22452/jati.vol25no1.8

Sharadgah, T. A., & Sa'di, R. A. (2022). A systematic review of research on the use of artificial intelligence in English language teaching and learning (2015–2021): What are the current effects? *Journal of Information Technology Education: Research, 21*, 337–377. https://doi.org/10.28945/4999

Shen, M. & Chiu, T. (2019). EFL learners' English speaking difficulties and strategy use. *Education and Linguistics Research, 5*(2), 88–102.

Stodden, R. A., Roberts, K. D., Takahashi, K., Park, H. J., & Stodden, N. J. (2012). Use of text-to-speech software to improve reading skills of high school struggling readers. *Procedia Computer Science, 14*, 359–362.

Suliman, W. F. N. (2020). *Investigating English speaking difficulties that encounter ESL learners* (Publication No. 33665535) [Masters' dissertation, The British University in Dubai]. ProQuest Dissertations & Theses Global.

Suro, F., & Ono, Y. (2016). Japanese EFL learners' uses of text-to-speech technology and their learning behaviors: A pilot study. In A. Hiramatsu, N. Komoda, T. Matsuo, & A. Kanzaki (eds.). *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 296-301). Los Alamitos, CA: IEEE Computer Society.

Tafazoli, D. (2024). Exploring the potential of generative AI in democratizing English language education. *Computers and Education: Artificial Intelligence, 7,* 1–11. https://doi.org/10.1016/j.caeai.2024.100275

Van Lieshout, C., & Cardoso, W. (2022). Google Translate as a tool for self-directed language learning. *Language Learning & Technology, 26*(1), 1–19.

Wardhani, W. M., Rafiqa, S., & Makmur, S. (2023). Problems faced by students in speaking English. *Borneo Journal of English Language Education, 5*(2), 1–9. https://doi.org/10.35334/bjele.v5i2.4683

Weir, C. J., & Wu, J. R. W. (2006). Establishing test form and individual task comparability: a case study of a semi-direct speaking test. *Language Testing, 23*(2), 167–197. https://doi.org/10.1191/0265532206lt326oa

Yahaya, B. S., Madzlan, N. A., & Muhammad, M. (2021). ESL learners' obstacles in learning speaking skills. *The English Teacher, 50*(3), 173–181. https://doi.org/10.52696/WYZY7207

Yang, H., & Kyun, S. (2022). The current research trend of artificial intelligence in language learning: A systematic empirical literature review from an activity theory perspective. *Australasian Journal of Educational Technology, 38*(5), 180–210. https://doi.org/10.14742/ajet.7492

Zhang, S. (2009). The role of input, interaction, and output in the development of oral fluency. *English Language Teaching, 2*(4), 91–100.

Appendix A

Semi-structured interview questions for the teacher

1. Study Implementation

- Can you describe how the study was carried out? What are your instructions to students?
- Can you describe how the pre-test and post-test were conducted?
- Did you encounter any challenges?

2. Student Engagement & Differences

- Comparing the two groups, what differences do you observe after introducing Natural Reader?
- Did any students resist using the tool? Why?
- Did you notice differences in how low/mid/high-level students interacted with Natural Reader?

3. Perceived Effectiveness

- Do you think Natural Reader is more useful for certain aspects of speaking (e.g., word stress, intonation)?
- Our preliminary analysis of scores shows that overall students' scores show marginal improvements, can you think of any possible reasons for this finding?
- Our preliminary analysis of scores also shows that while students' scores improved in general, only low-level students improved significantly, can you think of any possible reasons for this finding?

Appendix B

Focus group interview questions for students

1. Do you find your English speaking test easy/okay/difficult? What is the hardest part?
2. When you use Natural Reader in class, do you find it useful? Do you like to learn speaking by following Natural Readers more or by following your teacher more? Why?
3. Do you use Natural Reader outside of class? Why or why not?

學生訪談問題

1. 你覺得你的英語口語測試容易/還可以/困難嗎?最難的部分是甚麼?
2. 你在課堂上使用Natural Reader時, 覺得它有用嗎?你更喜歡跟著Natural Reader學習口語, 還是更喜歡跟著老師學習?為甚麼?
3. 你有在課外使用Natural Reader嗎?為甚麼有或為甚麼沒有?