PAEDAGOGIA
Jurnal Penelitian Pendidikan
http://jurnal.uns.ac.id/paedagogia

DOI : 10.20961/Paedagogia.v29i3.123190
Vol. 29, No. 3, 2026, pp. 546-562
p-ISSN 0126-4109
e-ISSN 2549-6670

Structured English Song Activities Versus Conventional Instruction in EFL Pronunciation Perception

Rosalina Heatubun, Mujiono*, Maria Cholifah
Department of English Education, Faculty of Languages and Arts, University of PGRI Kanjuruhan, Malang, Indonesia.

Keywords: *EFL learners; English Song Activities; Pronunciation Perception; Segmental and Suprasegmental Features; Quasi-Experimental Design*

Article history
Received: 5 July 2026
Revised: 11 September 2026
Accepted: 20 October 2026
Published: 31 October 2026

*Corresponding Author Email:
moejie_nova@unikama.ac.id

DOI: 10.20961/paedagogia.v29i3.123190

© 2026 The Authors. This open-access article is distributed under a CC BY-SA 4.0 DEED License

Abstract: Pronunciation remains a persistent challenge for EFL learners, particularly where exposure to authentic spoken English is limited. Although English songs are often used as phonological input, little research has examined their effectiveness in developing pronunciation through systematic practice of both segmental and suprasegmental features. This study investigated the effectiveness of structured English song activities for EFL students' pronunciation perception. A quasi-experimental pretest–posttest control group design was employed with 60 Grade VIII students from a public junior high school in Malang, Indonesia, drawn from two intact classes: an experimental group (n = 30) and a control group (n = 30). Over six weeks, the experimental group took part in structured song-based activities. In contrast, the control group received conventional instruction using textbooks and non-song materials with equivalent objectives and duration. Learning was measured with a multiple-choice listening test of auditory recognition and phonological discrimination rather than direct oral performance, so the findings concern perception-based pronunciation knowledge only. Gains were compared using change scores, an exploratory ANCOVA, and a non-parametric test. The two classes were not equivalent at baseline, as the experimental class scored higher at pretest. The findings showed no statistically significant improvement after the intervention and no significant difference between the groups. These results suggest that structured English song activities alone may be insufficient to produce measurable gains in classroom-based pronunciation practice. However, songs may still serve as complementary phonological input when combined with explicit modelling, guided repetition, individualized practice, and corrective feedback.

How to cite: Heatubun, R., Mujiono, & Cholifah, M. (2026). Structured English Song Activities Versus Conventional Instruction in EFL Pronunciation Perception. *PAEDAGOGIA, 29*(3), 546-562. DOI: 10.20961/paedagogia.v29i3.123190

INTRODUCTION

Teaching pronunciation is a highly problematic area of instruction in the context of EFL. Most of the teaching focuses on teaching vocabulary and grammar rather than pronunciation (Derwing & Munro, 2005; Mora, 2025). As a result of this omission, EFL learners usually cannot fully hear, distinguish, and practice segmental and suprasegmental features of English (Derwing & Munro, 2005; Mora, 2025). In addition to other features of English pronunciation, EFL learners must differentiate the features of spoken English and English orthography from the features of their L1 (Derwing & Munro, 2005; Mora, 2025). Features of speech of most importance include: (a) segments: vowel and consonant sounds; and (b) suprasegments: stress, rhythm, and intonation (Derwing & Munro, 2005; Mora, 2025). Segments differ from suprasegments, which apply to longer, interconnected units of speech. EFL learners must also know the English stress system, rhythm system, and intonational systems (Hodgetts, 2020; Yenkimaleki & Heuven, 2021).

Different strategies have been recommended to deal with pronunciation difficulties, such as direct explanation of sound-to-spelling rules, auditory discrimination practice, use of communicative games, and the correction of pronunciation errors. Research suggests that practicing targeted speech combined with models of expected behavior will lead to improvements in pronunciation over time (Ma et al., 2024; Wang, 2022; Yenkimaleki et al., 2023; Yenkimaleki & Heuven, 2021). The instruction of pronunciation should employ practices that allow learners to put what has been taught into a context with

higher demands, as opposed to doing a repetitive drill of individual pronunciation elements (Gordon, 2021; Mora, 2025; Yenkimaleki et al., 2023). Although there is a desire to improve this instruction, there are some roadblocks to implementing this recommendation in EFL settings. Because of a multitude of reasons, the majority of EFL instruction programs have heavily concentrated on monotonous and repetitious drilling practices, which provide little, if any, pronunciation or communicative opportunities for learners. Furthermore, stress is rarely used in speech, creating an overall unfavorable environment (Deng, 2026; Perez, 2022). With these issues in mind, affordable, accessible and, more importantly, adaptable material to teach EFL is sorely needed.

Many English learners find singing along with popular English songs very useful for their pronunciation, as well as rhythm and pitch. Since singing requires learners to recognize certain English phonemes and produce specific linguistic patterns, it is expected that the songs used provide ample practice with those patterns. Likely, Li & DeKeyser, (2021) and Muñoz & Cadierno, (2021), as well as many other second language acquisition researchers, and many other phonologists would agree with the premise that, with songs as guided pronunciation activities, learners would gain the chance to practice sounds that make up words as well as more complete sentence structures. The purpose behind teaching a song as part of a pronunciation class is what makes singing an effective classroom pronunciation tool. As with all things used for the teaching of languages, singing, to be successful, must be intentional, as there is still plenty of entertainment value in singing, making songs ideal tools for motivation.

However, as mentioned, studies that examine song-based learning as a tool for English pronunciation instruction differ extensively in both learners, procedures, duration, targeted sounds, and what the learners must do to succeed (Baills et al., 2021; Sun et al., 2025; Zhang et al., 2023). The many instructional and testing approaches to pronunciation have led to difficulties in drawing comparisons across studies. Research on this area of instruction has rarely addressed the difference between pronunciation outcomes assessed via recognizing the sounds and the outcomes assessed by directly oralizing those sounds. Finally, most studies do not analyze both the segmental and the suprasegmental features in the context of the same controlled intervention study. Therefore, studies comparing the recognition of pronunciation features between auditory pronunciation instruction and other forms of instructional techniques, including song-based pronunciation instruction, are still lacking.

The study designed and executed here was built on the premise that repeated exposures to pronunciation can direct attention toward and provide practicing opportunities for auditory patterns of pronunciation when learners have some conscious reason for attention toward pronunciation features. For songs, increased exposure can provide the learning benefits of both the rhythm and repetition that accompany the words of the songs. Still, learning pronunciation is assumed to depend more on focused attention and guided instruction/practice and pronunciation feedback rather than mere song exposures (Saito, 2021; Sun et al., 2025; Yenkimaleki & Heuven, 2021). Consequently, the song exercises used in the current study were used to fulfill many instructional needs of pronunciation. Along with being motivating, the exercises were incorporated into structured listening, pronunciation observation, guided pronunciation, and feedback stages. This helped structure the exposure to pronunciation while ensuring there were opportunities for directed practice of the targeted pronunciation features.

The focus of the present research was the effect of combining singing, exercises based on phonological properties, and feedback with structured activities on junior high EFL students' auditory recognition and phonological discrimination of segmental and suprasegmental features. This led to the creation of the research question:

RQ1. Does structured English song activity significantly improve EFL learners' overall pronunciation practice compared with conventional pronunciation instruction?

In addition, the current study created song-based exercises for teaching pronunciation and compared them with more traditional approaches. Junior high students' perception-based pronunciation knowledge, namely auditory recognition and discrimination of segmental and suprasegmental features, was the object of analysis. Accordingly, the objective of this study was to examine whether structured English song activities, compared with conventional pronunciation instruction, improved junior high EFL students' auditory recognition and phonological discrimination of segmental and suprasegmental

features over six weeks. The study is expected to contribute classroom-based evidence on song-based pronunciation input assessed through perception rather than production, and to inform whether songs are better used as a stand-alone activity or as a complementary resource.

METHOD

This research applied a quasi-experimental, pre-post design with non-equivalent control groups in two eighth-grade EFL classes to examine the effectiveness of singing songs with structured lyrics on improving phonology relative to traditional pronunciation instruction (D. T. Campbell & Stanley, 1963; Miller et al., 2020). There was no assignment of participants at random, but rather the intact class was either an experimental group or a control group (Shadish et al., 2002). Both groups were taught by the same teacher. Both groups experienced similar lengths of instruction, learned similar phonological sounds, and took similar tests at the same time.

Song activities in the experimental condition included the guided listening, notice-focused listening, lyric practice, guided practice with some feedback, and so on. Control class activities utilized traditional, scripted activities in the text, model audios, and guided practices with textbook-based audio. The objective of the activities in both classes was auditory phoneme recognition rather than actual articulation. Each of the treatment conditions had one intact class so that any treatment effect could be a function of the treatment or the class; consequently, the design only offered evidence for an impact of instruction with song-based audio over the standard pronunciation practice. Song-based materials were overly used in the experimental setting, as shown in Figure 1.

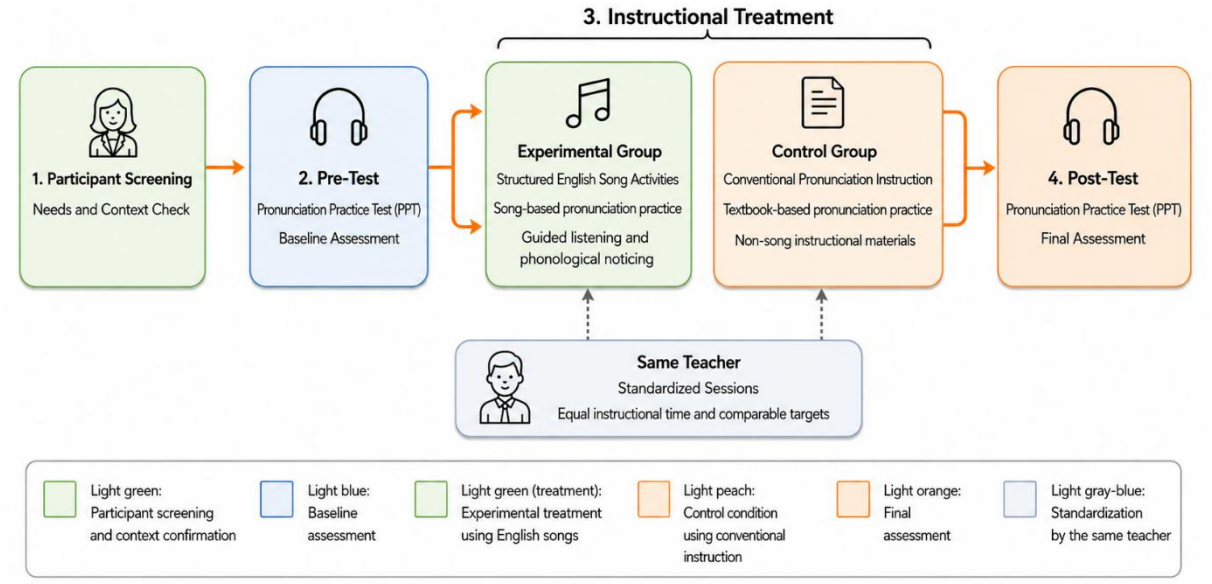


Figure 1. Quasi-Experimental Research Design

Research Design and Variables

The independent variable had two instructional conditions: Structured English Song activities versus conventional pronunciation instruction. The dependent variable of interest was the perception-based pronunciation score. This score reflected perceptual skills in discriminating and categorizing, and recalling features - mostly auditory skills. Features considered at the segmental level included contrastive vowel and consonant phonemes. The suprasegmental aspects of word stress, rhythmical phrasing, as well as intonation and sentence stress features, were taken into account. Direct pronunciation accuracy/comprehensibility and accentedness/intelligibility were not measured in this study. This definition of outcomes is in line with the purpose of this study, as stated in the introduction, and therefore limits the interpretation of the results to perception, recognition, and discrimination features.

Participants, Subject, and Group Assignment

In this study, 60 students of class VIII who were junior high school students in two classes in Malang, aged 13-14, participated. The classes were chosen as a matter of convenience because random assignment of classes was not feasible, since the classes were already arranged. A second consideration was school administration policy about the placement of students in class sections, which suggested the possibility that the two groups had not been equivalent prior to the intervention. Therefore, when interpreting the outcomes, the scientists considered this as a possibility.

The six-week segmental (and suprasegmental) course was the same for both classes, and the same instructor delivered it. The class that engaged in the experimental activities did song activities focusing on structured lyric exercises, phonological awareness activities, and song listening exercises, as well as focused practice and structured repetition with the instructor's guidance. The class that performed the traditional activity used textbooks and audio-based activities without songs, listening, and controlled repetition following the teacher's instructions. Only auditory recognition and discrimination were assessed, and the design included songs to provide more opportunities for focused listening and guided repetition of target features. The teacher had the same objective skills to teach for both groups and the same time to teach them.

This study was approved by the local school board, and then explained to the students, who gave their permission to participate. The participation was entirely voluntary, and the responses of all of these students were anonymously analyzed.

Instructional Treatment and Control Condition

The 12-session intervention lasted for six weeks, with two meetings per week of about 35-40 minutes each. For the experimental group, six pre-selected English songs formed part of their structured phonological input. In each experimental session, based on a song for the week, the target feature (for example, sound contrast, word stress, rhythm, and intonation) was introduced. For the sessions, a four-stage process was followed: pre-listening, guided listening, focused practice, and reflection. For pre-listening, the teacher introduced the target phonological feature and session goal. Learners were then guided to listen to the song and focus on the target phonological feature. For the focused practice session, the teacher engaged learners in repeated practice and activities based on the lyrics, while providing feedback on pronunciation. Each session closed with a reflection during which the target feature was reviewed briefly.

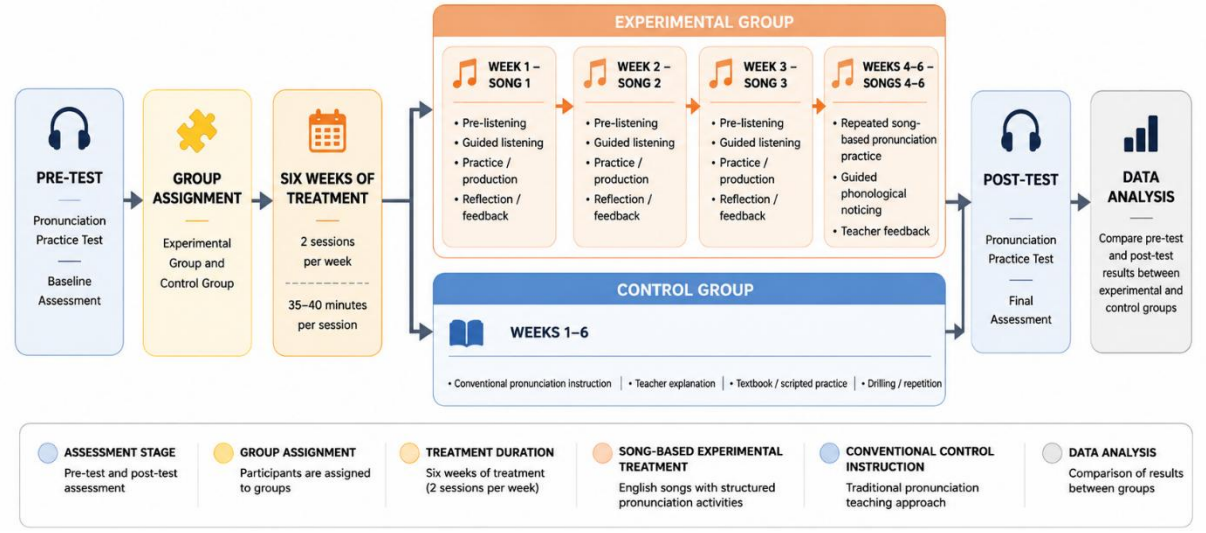


Figure 2. Research Timeline and Session Structure of the Six-Week Treatment in the Experimental and Control Conditions

The control group also practiced pronunciation using the same features and for similar times, but under a conventional format, again taught by the same teacher. This format used textbook exercises, listened to phrases, teacher demonstrations, controlled repetitions and non-song audios. The practice of features by control learners was driven by teacher demonstration, listening, repetition, and controlled

exercises. The songs and song-based lyrics used with experimental learners were not included. Thus, the major difference between the two groups was the type of instruction and format of practice for the same pronunciation features.

The PPT was also employed as an auditory and phonological test. No assessment of pronunciation production was undertaken. More details of each of the instruments used in the study are discussed in other sections of the article, as are details on the distributions, scoring, and reliability of the assessments. A pre-assessment was administered before the 6-week study began, and a post-assessment was administered afterward, as shown in Figure 2.

Selection of the Six Songs

A series of six English-language songs was selected based on the structured input needed for the research study. The songs met all pedagogical and linguistic requirements. They included such things as moderate speed of articulation, repetitive lyrical phrasing, clear and salient features, and accessible lyrics with little use of vocabulary, which are developmentally and chronologically appropriate for the audience, along with clear and loud sound quality. All criteria met eliminated any unnecessary difficulty in the use of listening and vocabulary, and offered more opportunities for learners to recognize patterns of sound as they were used in the classroom. This is as opposed to selecting songs on learners' preferred lists. It would have meant using a song of little pedagogical benefit. These selected criteria are used for classroom purposes and are explained in detail in Table 1.

Table 1. Criteria for Selecting the Six English Songs

Criterion	Purpose
Clear articulation	To help learners perceive pronunciation features more clearly
Moderate tempo	To reduce listening difficulty and support phonological noticing
Repetition in lyrics	To increase exposure to target sound patterns
Segmental and suprasegmental salience	To align the songs with instructional pronunciation targets
Age-appropriate content	To ensure suitability for the learners' classroom context
Lexical accessibility	To avoid vocabulary overload that may distract learners' attention
Good audio quality	To maintain clarity during classroom playback

Research Instrument, Domains, and Scoring

Pronunciation Practice Test (PPT) was the major tool used for the assessment of auditory pronunciation recognition and phonological discrimination. In line with the perceptual-production paradigm as laid out by Derwing & Munro (2005) and Isaacs & Trofimovich (2012). The PPT asks users to use perception to segment and categorize the pronunciation features from acoustical stimuli.

Table 2. Blueprint of the Pronunciation Practice Test

No	Domain	Indicator	Description of Practice Focus	Item Format
1	Segmental	Vowel contrasts	Identifying differences in English vowel sounds is commonly problematic for EFL learners.	Multiple choice
2	Segmental	Consonant contrasts	Distinguishing consonant sounds and sound substitutions.	Multiple choice
3	Suprasegmental	Word stress	Identifying correct stress placement in words	Multiple choice
4	Suprasegmental	Rhythm	Recognizing appropriate rhythm or syllable timing patterns.	Multiple choice
5	Suprasegmental	Intonation	Identifying suitable intonation patterns in short spoken contexts.	Multiple choice

The test had 30 questions in all: 15 in each of the segments and suprasegments. Vocal and consonant features were included under segments, and stress, stress timing and difference in intonation patterns were addressed under suprasegments. Given the focus of the study on recognition and discrimination, the format of the test was multiple-choice listening. The pretest was administered in Form A, and the posttest was administered in Form B. The test structure and scoring were the same for both tests. They were also given in the same manner to ensure consistency. More details about the different categories and test questions are included in Tables 2 and 3.

Table 3. Blueprint for Item Distribution

No	Domain	Subdomain	Recommended Number of Items	Percentage
1	Segmental	Vowels	8	26.70%
2	Segmental	Consonants	7	23.30%
3	Suprasegmental	Word Stress	5	16.70%
4	Suprasegmental	Rhythm	5	16.70%
5	Suprasegmental	Intonation	5	16.70%
	Total		30 items	100%

Correct answers were scored as 1 and incorrect answers as 0. During item analysis, three items (Items 2, 19, and 30) did not meet the retention criteria, leaving 27 retained items (14 segmental and 13 suprasegmental). The scores analyzed in this article are percentages computed from all 30 administered items (correct responses / 30 × 100) and were not recomputed for the retained items. Table 4 presents the retained-item scoring procedure, which applies once item-level responses are rescored.

Table 4. Scoring Procedure of the Pronunciation Practice Test

No	Component	Scoring Rule
1	Correct answer	1 point
2	Incorrect answer	0 point
3	Segmental score	Sum of vowel and consonant items
4	Suprasegmental score	Sum of stress, rhythm, and intonation items
5	Overall score	Sum of all items

Instrument Validity and Reliability

The PPT items were judged for content validity by a panel of EFL instructors based on four criteria, namely accuracy, clarity, relevance, and level of instruction. A sub-group of items was then agreed for inclusion in the target population and in the learning outcomes after considering comments raised by the various panels.

Three of the 30 test items (Items 2, 19 and 30) were not retained in the item analysis due to the items' failure to meet the retention criteria. By dichotomously scored items, KR-20 is an internal consistency estimate that is comparable to Cronbach's alpha. Cronbach's alpha for the 30-item test, before item exclusion, was .671 for the full test and .585 for the segmental component, with no estimate for the suprasegmental component. The 27 items that were retained were not separately calculated for Form A and Form B, and these coefficients were not recalculated. The blueprint and scoring structure were the same for the two forms; content and difficulty have not been empirically shown to be equivalent across forms.

Data Collection

Data collection included pretest, intervention, and posttest data. Both groups completed the PPT form A in the pretest, the same as the post test, and under the same testing circumstances, both in terms of timing, content, instructions, and scoring procedures. The experimental group underwent 12 English-song lessons during the intervention, which spanned 6 weeks. Participants in the control group had 12 sessions of more traditional pronunciation lessons during the same time period. Despite differences in teaching materials, both groups highlighted teaching of segmental and suprasegmental features in their lessons.

Form B of PPT was administered to both groups of participants at the end of the 6 weeks under

the same conditions as the pretest, Form A. The use of different forms decreased the chance that the participants would remember specific pretest items. The blueprint, scoring, and administration of both tests were the same.

Data Analysis

The analyses were done by every participant's percent score on the Pronunciation Practice Test (PPT). Each class (control and experimental) mean, standard deviation, minimum, and maximum were computed at pretest and posttest, along with the change from pretest to posttest in each group. As there was only one whole class for each condition, the results pertain to the two observed classes.

Normality of the distributional assumptions was checked using Shapiro–Wilk tests on the distribution of raw scores and Levene's test for equality of variances. The difference in change between the two classes was primarily compared using the difference in mean change (experimental minus control), and an independent-samples t-test was conducted on the student level change scores, with Cohen's d used to calculate the standardized effect size. Given the non-random assignment of treatments and assignment of the unit of treatment (the intact class), p-values were used as a descriptive tool rather than a confirmatory test of a treatment effect.

Instructional condition was the independent variable, pretest score was the covariate, and posttest score was the dependent variable for an exploratory student-level ANCOVA. Partial η^2 is reported as the effect size. Multilevel modeling was not possible due to nesting students in classes, and there were only two classes available, making it impossible to separate the instructional condition from class membership in the ANCOVA. No pretest-by-condition interaction and residual diagnostics are provided. The posttest scores distribution for the experimental class was not normal, and a Mann–Whitney U test was used as a sensitivity check. Still, it does not correct for baseline differences or the clustering of students within classes.

RESULT

This section presents the item screening and reliability evidence for the Pronunciation Practice Test (PPT), the descriptive pretest and posttest results, the distributional diagnostics, and the comparisons between the two classes. All outcomes concern auditory recognition and phonological discrimination of segmental and suprasegmental features; productive pronunciation was not assessed. Because each condition was represented by a single intact class, the results reflect the two observed classes and should be interpreted with caution.

Item Screening and Reliability

Of the 30 PPT items administered, 27 met the retention criteria (90.00%). In the segmental domain, 14 of 15 items were retained, and Item 2 was excluded. In the suprasegmental domain, 13 of 15 items were retained, and Items 19 and 30 were excluded. The retention rates were therefore 93.33% for the segmental domain and 86.67% for the suprasegmental domain (Table 5). Cronbach's alpha was .585 for the segmental component and .671 for the full 30-item test; a separate coefficient for the suprasegmental component was not available. These coefficients were computed before item exclusion, were not re-estimated for the 27 retained items or separately for Forms A and B, and are therefore provisional.

Table 5. Item Screening and Reliability of the Pronunciation Practice Test

Domain	Administered Items	Retained Items	Invalid Items	Retention Rate	α
Segmental	15	14	2	93.33%	.585
Suprasegmental	15	13	19, 30	86.67%	Not available
Overall	30	27	2, 19, 30	90.00%	.671

Note. Coefficients were computed before item exclusion and have not been recalculated for the 27 retained items.

The results presented in Table 5 indicate a slight difference between the segmental and

suprasegmental domains in item retention. Both reliability coefficients are below the .70 alpha level that is conventionally used, and the segmental coefficient is the lower of the two. There was a lack of internal consistency, which may have affected the test's sensitivity to minor changes. The PPT should be considered, therefore, as a short-duration measure of auditory recognition and discrimination, not as a precise measure of individual ability.

Descriptive Performance Across Pretest and Posttest

Table 6 presents the pretest and posttest scores of the two classes.

Table 6. Descriptive Statistics of Pretest and Posttest Scores

Test	Group	N	Mean	Std. Deviation	Minimum	Maximum
Pre-test	Control Group	30	56.67	16.31	23.33	83.33
Pre-test	Experimental Group	30	64.78	21.02	26.67	100.00
Post-test	Control Group	30	60.89	16.86	30.00	100.00
Post-test	Experimental Group	30	64.78	21.11	30.00	100.00

Note. Scores are percentages of all 30 administered items and were not recomputed for the 27 retained items.

The two classes were not the same at the beginning. At pretest, the experimental class (M = 64.78, SD = 21.02) scored 8.11 points higher than the control class (M = 56.67, SD = 16.31). At posttest, the experimental class remained at M = 64.78 (SD = 21.11), whereas the control class rose to M = 60.89 (SD = 16.86), narrowing the gap to 3.89 points. There was considerable variation in recognition performance within each class, as the standard deviations were large and individuals' scores ranged from 23.33 to 100.00 across the two measurement occasions.

Table 7. Pretest-to-Posttest Change in Mean Scores

Class	Pretest Mean	Posttest Mean	Mean Change	Relative Change	Interpretation
Control	56.67	60.89	+4.22	+7.45%	Small descriptive increase
Experimental	64.78	64.78	0.00	0.00%	No descriptive mean change
Experimental – Control	+8.11 at baseline	+3.89 at posttest	−4.22 difference in change	—	Baseline gap narrowed

Note. Relative change is calculated from the class means. Between-class values are experimental minus control.

Results in Table 7 reveal that the experimental class mean score was 4.22 points higher (+7.45% relative to the pretest mean) than the control class mean score, which showed no change. The mean change difference was thus -4.22 points (experimental - control). Since instructional condition was a confound of the class condition, this pattern was observed in both classes and is not indicative of the effectiveness of conventional instruction relative to song-based instruction.

Distributional Diagnostics

Shapiro–Wilk tests indicated that the control-class pretest, experimental-class pretest, and control-class posttest distributions did not depart significantly from normality (p = .135, .232, and .394, respectively). The experimental-class posttest distribution did depart from normality (W = .890, p = .005; Table 8).

ANCOVA assumes normally distributed residuals rather than normally distributed raw scores within groups, so the non-normal experimental posttest distribution does not by itself invalidate the ANCOVA; however, residuals and possible outliers were not examined in the reported output. Levene's tests, reported in Table 9, showed no significant evidence of unequal variances at pretest (p = .182) or posttest (p = .283).

Table 8. Shapiro-Wilk Tests of Raw Score Distributions

Measurement	Class	W	df	p	Decision
Pretest	Control	.946	30	.135	No significant departure from normality
Pretest	Experimental	.955	30	.232	No significant departure from normality
Posttest	Control	.964	30	.394	No significant departure from normality
Posttest	Experimental	.890	30	.005	Significant departure from normality

Note. Tests were applied to raw score distributions. Residual diagnostics and homogeneity of regression slopes are not reported.

Table 9. Levene's Test for Homogeneity of Variance

Measurement	Levene Statistic	df1	df2	p	Interpretation
Pretest	1.828	1	58	.182	No evidence of unequal variances
Posttest	1.173	1	58	.283	No evidence of unequal variances

Note. Equal variances do not resolve the class-level confounding created by one intact class per condition.

Table 9 indicates comparable variances across the two classes at both time points. This supports homogeneity of variance but not the remaining ANCOVA assumptions, and it does not address the confounding of condition with class membership.

Between-Class Comparison of Change

The primary comparison concerns the difference in change between the two intact classes. The control class improved by 4.22 points, whereas the experimental class showed no change, a difference of −4.22 points (experimental minus control; Table 10). A student-level comparison of change scores was not statistically significant, $t(58) = 0.652$, $p = .517$, and the standardized difference was small (Cohen's $d = .168$, in the direction of the control class's larger gain). Because the unit of assignment was the intact class, this p-value should not be read as a confirmatory test of a population-level effect. The data show no practically meaningful advantage of the song-based class on the PPT.

Table 10. Between-Class Comparison of Change

Statistic	Value	Interpretation
Control mean change	+4.22 points	Descriptive increase
Experimental mean change	0.00 points	No descriptive mean change
Difference in change (Experimental − Control)	−4.22 points	Change favored control descriptively
Change-score t-test	$t(58) = 0.652$, $p = .517$	Not statistically significant in the reported student-level analysis
Cohen's d	.168	Small standardized difference

Note. The inferential test is reported for transparency and should be interpreted with caution because students were nested within a single class per condition.

Table 10 shows that the control class's mean gain exceeded that of the experimental class by 4.22 points, but the difference was small and not statistically significant. It describes the two observed classes and cannot be attributed to the instructional method, because students were nested within a single class per condition.

Exploratory Baseline-Adjusted ANCOVA

As an exploratory analysis, posttest scores were compared across conditions with pretest scores as the covariate. The effect of instructional condition was not significant, $F(1, 57) = 0.320$, $p = .574$, partial $\eta^2 = .006$, and neither was the pretest covariate, $F(1, 57) = 0.904$, $p = .346$, partial $\eta^2 = .016$ (Table 11). The analysis is exploratory because the instructional condition cannot be separated from class membership, and the pretest-by-condition interaction and residual diagnostics are not reported.

Table 11. Exploratory ANCOVA for Posttest Scores

Effect	F	df	p	Partial η²	Interpretation
Pretest covariate	0.904	1, 57	.346	.016	Small; not statistically significant
Instructional condition	0.320	1, 57	.574	.006	Very small; not statistically significant

Note. The ANCOVA is exploratory because only one intact class represented each condition.

After adjustment for pretest scores, the song-based and conventional classes did not differ significantly at posttest, and the effect sizes were very small. This is consistent with the change-score comparison; both analyses are limited by the use of a single class per condition.

Sensitivity Analysis

As a sensitivity check, a Mann–Whitney U test was applied to the posttest scores because the experimental-class distribution departed from normality. The difference was not significant, U = 419.50, z = −0.445, p = .656, r = .058 (Table 12), indicating a negligible effect. Like the other analyses, this test does not adjust for baseline differences or for the clustering of students within classes.

Table 12. Mann-Whitney U Test of Posttest Scores

Analysis	Statistic	p	Effect Size	Interpretation
Posttest comparison	U = 419.50; z = −0.445	.656	r = .058	Very small; not significant

Note. r = effect size.

Table 12 shows that the posttest score distributions of the two classes did not differ significantly. This result agrees with the change-score and ANCOVA analyses and should be read as supporting evidence only.

Answer to the Research Question

Regarding RQ1, the results provide no evidence that structured English song activities improved junior high EFL students' auditory recognition and phonological discrimination more than conventional pronunciation instruction. Although the song-based class scored higher at baseline, its mean did not change over six weeks, whereas the conventional class gained 4.22 points on average. The difference in change was small and not significant in the change-score comparison (p = .517), the exploratory ANCOVA (p = .574), and the Mann–Whitney test (p = .656). These findings do not show that songs are ineffective for speaking, because the PPT measured recognition rather than production. Nor do they establish the equivalence or superiority of either approach, because each condition was represented by one intact class.

DISCUSSION

Results are provided for the perception-based pronunciation outcome measure taken at 6 weeks, revealing no observable benefits of structured English song activities over traditional instruction. This discovery calls into question the statement that a certain amount of musical input in the teaching of pronunciation leads automatically to more phonological learning. The overall results of the present classroom comparison did not indicate a similar advantage in terms of auditory recognition and discrimination. Pronunciation benefits under specific training conditions (Baills et al., 2021; Zhang et al., 2023). These outcomes should therefore be interpreted as comparisons between evidence of learning and assessment in different learning and assessment arrangements, not as direct comparisons of different estimates of the same action in the same instructional context. Although the relationship between music and second language learning was generally positive, there were certain inconsistent findings regarding the statistical support in the literature Sun et al. (2025). A study also found methodological flaws in song-based intervention research in such a way that definite causal claims of linguistic benefits cannot be made (Hamilton et al., 2024). All these reviews work together to allow for a conditional interpretation of musical input rather than an expectation of a uniformity of effectiveness. Modelling, listening, repetition

and guided practice were included in both the conditions of this current study, and the instructional comparison was between the added value of the song format. The present pattern, thus, does not prove that songs are not of educational significance. Rather, it suggests that their additional contribution to the observed receptive outcome was not evident in these two classes. The study thus further widens the scope of the debate on song-based instruction, establishing a classroom context in which its supposed phonological benefit has not been supported.

There was no overall mean improvement on the recognition test administered as a result of the intervention in the experimental class. Although there are differences in the training tasks and outcome measures, this pattern deviates from the positive results of pronunciation produced by Bailles et al., (2021) and Zhang et al., (2023). The repetition of music might have enhanced recognition of specific lyrics without leading to enough generalized recognition of the sampled features in the posttest. This is only a hypothesis, and the study has not assessed familiarity and transfer separately. Another study, the updated Speech Learning Model, offers a pertinent theoretical framework, as it ties the development of phonetic categories to the amount and quality of input to that category, as well as to first language (L1) categories (Flege & Bohn, 2021). In that context, repeated exposure may be useful, but repetition is not evidence of phonetic reorganization. Comparable results were reported by Uchihara et al., (2025) in their synthesis of high variability phonetic training, which showed significant perceptual learning in training conditions with multiple talkers, multiple phonetic contexts, and multiple types of corrective feedback. The current activities did not have that training paradigm documented, and therefore, the advantages of learning that way cannot be assumed to transfer to the song condition. This comparison implies that it is not the repetition of musical material per se, but the organization of perceptual practice that is more important. However, a no change means does not say anything about the fact that no students changed. Specific enhancements and declines may have compensated for each other, thus preventing the overall trend from exposing specific learning trajectories.

The descriptive improvement in the control class suggests that conventional instruction served as an active pedagogical comparison that provided phonological opportunities. Both groups were given modelling and guided repetition, and so the extra value of the songs lay further than the already existing instructional elements in the control. Positive outcomes of pronunciation instruction, and feedback duration and level of intervention were among the moderators of the reported outcomes J. Lee et al. (2015). Additionally, B. Lee et al., (2020) proved the importance of differentiating between perception-based and production-based practice. Furthermore, Yenkimaleki & Heuven (2021) also found that phonological focus and practice orientation are important when decoding pronunciation instruction. Together, these studies suggest that the organization of practice is a more powerful account of practice than a dichotomy of musical and nonmusical materials. There is also an additional justification for studying the question of what dimensions of pronunciation instruction are truly aimed at, as offered by Saito (2021). The authors suggest that the rise of the control class might be due to the spoken models in the recognition test being more similar to those used by the control class. Materials available, however, do not provide the acoustic relationship between instructional recordings and assessment stimuli. However, it may be that there were baseline differences, involvement in the classroom and practice that could not be measured outside of lessons that were also a factor in the observed change. When considering how the control class improved, this should be understood as a description of the classroom path, not as evidence of the causal superiority of conventional instruction.

The higher average score for the experimental class reflects differences that existed prior to the intervention and makes the interpretation of learning trajectories generated later more difficult. While the higher mean of the posttest is an advantage of the song intervention, it was reduced from 8.11 points to 3.89 points. However, the higher descriptive gain of the control class cannot be attributed solely to traditional teaching. The potential for more observable improvement with a lower initial mean was not specifically explored. Another possible explanation is the selection effect (also called regression to the mean), but this cannot be quantified from the available summaries. The current evidence thus suggests a comparatively more limited inference: Song learning is not necessarily more favorable in a genus of songbirds. A methodological context for this research is given by Hamilton et al. (2024), who identified limitations in the causal evidence for classroom song interventions. Transparent methods and defensible

measurement are also important, according to Kostromitina et al. (2025), in the study of pronunciation. In the current design, there was only one intact class for each condition, and the treatment was confounded with class assignment. This confounding is minimized by using the same teacher and the same time, but not eliminated. The shrinking gap is thus an indicator of knowledge about the two classes observed, but not enough to determine an external instructional impact.

The nature of the assessment limits the conclusions of the study to only auditory recognition and phonological discrimination, and not spoken pronunciation development. Understanding a good contrast is not necessarily the ability to express it correctly in interaction. Likewise, the realization of a stress position does not show a proper stress realization in the context of connected discourse. Studies B. Lee et al. (2020) and Nagle & Baese-Berk (2022) give reasons to consider perception and production to be related but distinguishable aspects of speech learning. In Flege & Bohn (2021), a theoretical description is provided in which the two modalities are not automatically processed in the same sequence but rather in an interacting fashion. Empirically, Uchihara et al. (2025) reinforce this distinction with their modest results in production advantage following perceptual training and the limited evidence for production generalization and continual production learning. Thus, the present receptive results cannot date back to nor refute oral changes in accuracy, intelligibility, comprehensibility or accentedness. Multiple-choice also allows for response demands that are not the same as those that are required when generating spontaneous speech. The potential benefits of practicing lyrics may then lie beyond the scope of the instrument. These advantages are unproven due to the lack of production samples for testing by the oral route. The study provides evidence for a specific receptive outcome, but the transfer to communicative pronunciation remains unresolved.

There are no segmental and suprasegmental learning outcomes reported separately, so a conclusion regarding similarity in response to the intervention in both can not be drawn. This instrument measured both domains, but the total score represents distinct patterns of change. An improvement in one element may have been counterbalanced by another element's poor performance. In either case, the components might not have changed much, but it is not possible to tell the difference with the information available. A study by Yenkimaleki & Heuven (2021) shows that the relationship between phonological focus and practice-orientedness deserves to be studied. A study by Saito (2021) also distinguishes segmental, prosodic, and temporal dimensions for interpreting pronunciation results. In song-based activities, the allocation of attention to sounds, stress, and intonation may logically vary with musical rhythm, but was not measured here. The measurement framework put forth by Saito & Plonsky (2019) can be useful for determining which dimension(s) each assessment addresses rather than viewing pronunciation as one monolithic construct. However, Uchihara et al. (2025) also demonstrate how target speech sounds and testing conditions can contribute to perceptual training results, with the main focus on phonetic training, not song-based prosodic training. These comparisons suggest that there is no need to assign superior or inferior status to the different domains; they merely indicate that further study of the domains may be warranted. Also, reliability evidence would need to be sufficient for use in separate analyses for each component retained. In the meantime, the combined score can only be used as an overall recognition measure.

The results of the reliability and item screening raise doubts about the instrument's ability to distinguish between small changes in instruction and measurement variability. The reported overall alpha of .671 and segmental alpha of .585 do not engender confidence in the precision of the instrument, particularly in the fact that they were not recalculated for the final retained item set. These coefficients by themselves do not render the test unusable. However, measurement inconsistencies may mask small learning differences. Reliability, validity and transparency are all major issues in pronunciation assessment, as suggested by Kostromitina et al. (2025) in their synthesis, which mainly covers a listener-based assessment of speech. A study by Saito & Plonsky (2019) also focuses on the interplay between the construct being assessed and how it is operationalized to understand instructional results. J. Lee et al. (2015) and Uchihara et al. (2025) also show that the instructional impact reported depends on aspects of the assessment. Scores for all 30 items administered were utilized in the analyses reported here, not the 27 items retained. Three items are removed from the test, and score distributions and comparative estimates may change. Still, the direction and magnitude of such changes cannot be determined without

rescoring the item-level responses. The comparability of the pretest and posttest forms is not only in the structure of the instruments. Interpretation is thus provisional until the items are rescored in the retained sample and the measurement properties of both forms are documented.

The reported comparative estimates of the small size do not indicate any observed support for a large song advantage, but do not indicate instructional equivalence. Student-level analyses of the change scores did not result in any statistically significant differences across treatments for the change-score comparison and exploratory ANCOVA. However, the lack of statistical significance does NOT imply that the population effect is zero. These standardized effect sizes were also small (Cohen's d of .168 favoring the control class, or r of .058 for the Mann–Whitney test) and are imprecise because of the presence of only one class per condition. A study by Kostromitina et al. (2025) offer a pertinent justification for clearly stating the methods and metrics used before making material inferences. A study by Hamilton et al. (2024) also warn against making strong causal inferences where inadequate methodological support is provided by the intervention design. The unanimous conclusion of several analyses can not remedy the fact that they are all unable to distinguish between the instructional condition and class membership. Also, the reported distributional tests do not support the assumptions of residual and regression slope required for the exploratory ANCOVA. Studies Nagle & Baese-Berk (2022) and Saito & Plonsky (2019) again emphasize the importance of linking the interpretation of statistical results with measurement and design, and not relying solely on the evidence of analytical convergence for that. The conclusion, as a result, is confined to the pattern observed in a specific classroom; effectiveness, ineffectiveness, and equivalence are still questions for which evidence is needed.

The range of scores is large, suggesting large differences between learners that the classroom means cannot account for. The standard deviation of scores for the experimental class of about 21 points indicates significant differences in recognition pre- and post-instruction. This variation is in relation to the auditory result measured and not as proof of a grammatical and/or orthographical difference. The theoretical groundwork for examining the role of previous phonetic categories and previous input in individual differences in speech learning is provided by Flege & Bohn (2021) and Nagle & Baese-Berk (2022) make the same point about the developmental complexity of perception–production relationships. Learners may also vary in their knowledge and experience of English songs and in their prior exposure to the chosen lyrics and opportunities to practise outside school. Because of the lack of a measurement of the relevant variables, these explanations are speculative. A study by Uchihara et al. (2025) demonstrates that learner profiles and training characteristics may affect the outcomes of phonetic learning, and this invites further research into the differences, which were not discussed in this sample. The spread thus suggests that an unchanged class mean might mask a variety of responses to teaching. A focus on each subject's change, in conjunction with an examination of the baseline proficiency level, would better clarify this possibility and avoid confusion with limited overall progress. Proposed moderators should be assessed in the future using prospective rather than retrospective assessment of score variability.

The integrated findings indicate that structured English songs could be a potentially beneficial instructional complement whose extra contribution in terms of phonology in the context of these songs has yet to be confirmed. This interpretation could be reconciled with the positive results obtained from Bailles et al. (2021) and Yenkimaleki et al. (2023); however, it does not imply that the gains realized in these studies are directly applicable to other classroom settings. It also aligns with the work of Hamilton et al., (2024) and Sun et al. (2025), who speak of the boundaries in the evidence of linguistic benefits and the importance of music in the context of the broader pedagogical approach. A logical extension is to link the chosen lyric to other spoken examples and recognition activities beyond just the songs practised. Individual responses should be checked – collective repetition does not mean that each learner has noted the feature(s) being worked on. When objectives are set for spoken communication, receptive practice should be enhanced by providing opportunities to produce the features to be learned in speaking. Studies J. Lee et al. (2015) argue for a focus on feedback, and Uchihara et al. (2025) give a justification for systematic variation of the phonetic input in perceptual practice. The following proposals are implications for instructional development, not necessarily mechanisms illustrated by the present data. Future investigations should investigate several classes in each condition, with consistent scoring, domain-specific outcomes, and delayed assessment of both recognition and production. They should also record

implementation in such a way that it shows a difference between musical format and the quantity and quality of practice. The contribution of the study was therefore to qualify the general hypotheses of instruction by song, suggesting that its added value must be demonstrated directly, and thus involves a sufficiently controlled design and sufficiently adapted measures.

Limitations

This study has a number of methodological drawbacks that need to be taken into account when interpreting the results. Firstly, the study was conducted with a quite small number of students (60 students in one school), so that the results of the study cannot be generalized to other EFL learners and other educational contexts. Second, participants were recruited from only two pre-existing classes (one class assigned to each treatment condition). All students in a condition were in the same class, which prevented any separation of the impact of the instructional condition from the possible impact of class-related factors such as the makeup of the class, classroom context, and other factors that were not measured. The statistical analyses were also conducted at the student level and thus did not replicate the treatment contrast at the level of the classes (J. P. Campbell, 2012).

The other major limitation is related to the instrument of assessment. The Pronunciation Practice Test (PPT) was multiple-choice and focused primarily on students' recognition and knowledge of the selected segmental and suprasegmental pronunciation features, rather than on how students produced pronunciation. Hence, the instrument was not able to give evidence directly on the accuracy of the students' production of the target pronunciation features. Furthermore, segmental and suprasegmental pronunciation production aspects were not measured separately. Due to these limitations, the results only bear witness to the recognition and knowledge indicated by the PPT in terms of pronunciation. Finally, scores were calculated for all 30 items administered. However, three of the items (2, 19, and 30) did not meet the retention criteria. The reliabilities (segmental component $\alpha = .585$ and overall $\alpha = .671$) were computed prior to item deletion and did not meet the traditional .70 criterion, potentially reducing the sensitivity of the test to small changes.

Recommendations

As a result of these restrictions, studies with larger and more diverse subjects, such as learners from different classes, schools or educational contexts, can be recommended in the future, so that findings can be generalized. Future studies should also employ more comprehensive assessment procedures that assess both segmental and suprasegmental pronunciation. In particular, the recognition-based or multiple-choice type of instruments should be supplemented by direct oral production instruments, such as read-aloud instruments, recorded oral production instruments, and other structured oral production instruments involving appropriate evaluation procedures.

Future research should also explore the use of different measures of segmental and suprasegmental features to investigate changes in individual pronunciation components more precisely. It would be better to use a combination of tests dealing with recognition, perception, and production to obtain a more comprehensive measure of pron. progress. It would be better to be able to say from these tests whether a structured English song activity has an impact on helping the learner recognize pronunciation features, produce them accurately, or both.

Lastly, the potential use of English songs as a secondary source of phonological input, rather than as a pronunciation intervention program in its own right, can be investigated. Research might explore activities involving music listening and noticing, explicit pronunciation, guided repetition, controlled production, and corrective feedback. This research would result in more compelling evidence of songs to help with various elements of pronunciation acquisition.

CONCLUSION

The findings of this study indicate that structured English song activities did not produce a statistically significant advantage over conventional pronunciation instruction in the overall auditory recognition and phonological discrimination outcomes measured by the Pronunciation Practice Test (PPT). The baseline-adjusted ANCOVA showed no statistically significant difference between the experimental and control groups after controlling for pretest scores. The change-score and Mann–

Whitney U sensitivity analyses produced consistent results.

The results indicate that in the context studied, pronunciation achievements in recognition-based tasks might not be greater when structured English song activities are the only intervention than with traditional pronunciation instruction. The results should also be considered in light of the limitations of the study, in which intact classes were used, a limited sample was employed, no direct measure of oral production was used, scores on all 30 items administered were used, and the internal consistency of the assessment instrument was limited.

On a theoretical level, the results provide evidence to support research on song-based pronunciation input in the classroom. The results did not demonstrate any perception-based gains from song activities over conventional instruction under the conditions of this study, calling into question the assumption that musical input alone leads to additional phonological learning. Teaching implications are that songs are not recommended as a sole pronunciation technique and should be presented as supplementary material to be followed by explicit modelling, guided repetition and corrective feedback in pedagogy. These inferences are provisional, as only recognition and discrimination were tested, and each condition was only sampled with one intact classroom.

Future studies can use larger and more varied samples, incorporate detailed measures of actual oral pronunciation production, and differentiate segmental and suprasegmental outcomes using appropriate evaluation procedures. A more complete picture of the potential role of structured English song activities in the various aspects of pronunciation development could be obtained from such designs.

AI USE STATEMENT

OpenAI's ChatGPT was utilized exclusively for grammatical revision and linguistic refinement to improve the clarity and accuracy of the manuscript. The tool was not used to generate or draft the manuscript's content, conduct data analysis, or interpret research findings. All aspects of the research, including data analysis, interpretation, and manuscript development, were undertaken independently by the authors, who assume full responsibility for the final publication.

REFERENCES

Baills, F., Zhang, Y., Cheng, Y., Bu, Y., & Prieto, P. (2021). Listening to Songs and Singing Benefitted Initial Stages of Second Language Pronunciation but Not Recall of Word Meaning. *Language Learning*, *71*(2), 369–413. https://doi.org/10.1111/lang.12442

Campbell, D. T., & Stanley, J. C. (1963). Experimental and Quasi-Experimental Designs for Research. In *Handbook Research on Teaching*. https://doi.org/10.1037/022808

Campbell, J. P. (2012). Behavior, Performance, and Effectiveness in the Twenty-first Century. In *The Oxford Handbook of Organizational Psychology* (Vol. 1). https://doi.org/10.1093/oxfordhb/9780199928309.013.0006

Deng, X. (2026). Working memory in technology-enhanced language learning: a systematic review from interactive to AI-mediated contexts. *Frontiers in Psychology*, 1–30. https://doi.org/10.3389/fpsyg.2026.1758104

Derwing, T. M., & Munro, M. J. (2005). Second Language Accent and Pronunciation Teaching: A Research-Based Approach. *TESOL Quarterly*, *39*(3), 379. https://doi.org/10.2307/3588486

Flege, J. E., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r). In *Second Language Speech Learning* (pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Gordon, J. (2021). Pronunciation and Task-Based Instruction: Effects of a Classroom Intervention. *RELC Journal*, *52*(1), 94–109. https://doi.org/10.1177/0033688220986919

Hamilton, C., Schulz, J., Chalmers, H., & Murphy, V. A. (2024). Investigating the substantive linguistic effects of using songs for teaching second or foreign languages to preschool, primary and secondary school learners: A systematic review of intervention research. *System*, *124*, 103350. https://doi.org/10.1016/j.system.2024.103350

Hodgetts, J. (2020). Suprasegmental Instruction and Intelligibility. In *Pronunciation Instruction in English*

for Academic Purposes (pp. 51–85). https://doi.org/10.1007/978-3-030-56116-1_3
Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition*, *34*(3), 475–505. https://doi.org/10.1017/S0272263112000150
Kostromitina, M., Sudina, E., & Baghlaf, E. (2025). Study and instrument quality in perception-based L2 pronunciation research. In *Studies in Second Language Acquisition* (Vol. 47, Issue 2). https://doi.org/10.1017/S027226312500018X
Lee, B., Plonsky, L., & Saito, K. (2020). The effects of perception- vs. production-based pronunciation instruction. *System*, *88*, 102185. https://doi.org/10.1016/j.system.2019.102185
Lee, J., Jang, J., & Plonsky, L. (2015). The Effectiveness of Second Language Pronunciation Instruction: A Meta-Analysis. *Applied Linguistics*, *36*(3), 345–366. https://doi.org/10.1093/applin/amu040
Li, S., & DeKeyser, R. (2021). Implicit Language Aptitude: Conceptualizing the Construct, Validating the Measures, and Examining the Evidence. *Studies in Second Language Acquisition*, *43*(3), 473–497. https://doi.org/10.1017/S0272263121000024
Ma, Q., Mei, F., & Qian, B. (2024). Exploring EFL students' pronunciation learning supported by corpus-based language pedagogy. *Computer Assisted Language Learning*, *0*(0), 1–27. https://doi.org/10.1080/09588221.2024.2432965
Miller, C. J., Smith, S. N., & Pugatch, M. (2020). Experimental and quasi-experimental designs in implementation research. *Psychiatry Research*, *283*, 112452. https://doi.org/10.1016/j.psychres.2019.06.027
Mora, J. C. (2025). Phonological acquisition in instructed SLA: L2 pronunciation training in and beyond the classroom and individual differences. *Language Teaching*, 1–16. https://doi.org/10.1017/S026144482510102X
Muñoz, C., & Cadierno, T. (2021). How do differences in exposure affect english language learning? A comparison of teenagers in two learning environments. *Studies in Second Language Learning and Teaching*, *11*(2), 185–212. https://doi.org/10.14746/ssllt.2021.11.2.2
Nagle, C. L., & Baese-Berk, M. M. (2022). Advancing the State of the Art in L2 Speech Perception-Production Research: Revisiting Theoretical Assumptions and Methodological Practices. *Studies in Second Language Acquisition*, *44*(2), 580–605. https://doi.org/10.1017/S0272263121000371
Perez, M. M. (2022). Second or foreign language learning through watching audio-visual input and the role of on-screen text. *Language Teaching*, *55*(2), 163–192. https://doi.org/10.1017/S0261444821000501
Saito, K. (2021). Effects of Corrective Feedback on Second Language Pronunciation Development. In *The Cambridge Handbook of Corrective Feedback in Second Language Learning and Teaching* (pp. 407–428). https://doi.org/10.1017/9781108589789.020
Saito, K., & Plonsky, L. (2019). Effects of Second Language Pronunciation Teaching Revisited: A Proposed Measurement Framework and Meta-Analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. In *Houghton, Mifflin and Company*. https://psycnet.apa.org/record/2002-17373-000
Sun, B., Yang, J., & Liang, Z. (2025). How music facilitates second language (L2) learning: a systematic review from 2002 to 2022. *Innovation in Language Learning and Teaching*, *19*(2), 155–173. https://doi.org/10.1080/17501229.2024.2350490
Uchihara, T., Karas, M., & Thomson, R. I. (2025). High variability phonetic training (HVPT): A meta-analysis of L2 perceptual training studies. In *Studies in Second Language Acquisition* (Vol. 47, Issue 3). https://doi.org/10.1017/S0272263125100879
Wang, X. (2022). Segmental versus Suprasegmental: Which One is More Important to Teach? *RELC Journal*, *53*(1), 194–202. https://doi.org/10.1177/0033688220925926
Yenkimaleki, M., & Heuven, V. J. van. (2021). Effects of attention to segmental vs suprasegmental features on the speech intelligibility and comprehensibility of the EFL learners targeting the perception or production-focused practice. *System*, *100*, 102557. https://doi.org/10.1016/j.system.2021.102557

Yenkimaleki, M., van Heuven, V. J., & Soodmand Afshar, H. (2023). The efficacy of segmental/suprasegmental vs. holistic pronunciation instruction on the development of listening comprehension skills by EFL learners. *Language Learning Journal*, *51*(6), 734–748. https://doi.org/10.1080/09571736.2022.2073382

Zhang, Y., Baills, F., & Prieto, P. (2023). Singing Songs Facilitates L2 Pronunciation and Vocabulary Learning: A Study with Chinese Adolescent ESL Learners. *Languages*, *8*(3), 1–24. https://doi.org/10.3390/languages8030219