International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

RESEARCH ARTICLE

Productive and Receptive Awareness of English Stress Following Explicit Phonology Instruction

Nadhim Aldubai[1] and Prof. Omer Elsheikh Hago Elmahdi[2]

1 *Department of Languages & Translation, Faculty of Arts and Humanities, Taibah University, Madinah, Saudi Arabia*
ORCID: https://orcid.org/0009-0005-3890-1300

2 *The Open University of Sudan, Affiliated to Department of Languages & Translation, College of Arts & Human Science, Taibah University, Saudi Arabia*
ORCID: https://orcid.org/0009-0009-1335-2607

Corresponding Author: Prof. Omer Elsheikh Hago Elmahdi, **E-mail:** Ohago65@gmail.com

ABSTRACT

This study examined changes in English stress production and perception among 26 Level 6 English majors at Taibah University, Saudi Arabia, assessed before and after eight weeks of explicit phonology instruction combining explanation, audio practice, spoken practice, and teacher feedback. A single-group repeated-assessment design was used. Production tasks involved words, sentences, and passages, and each recording was scored independently by two raters, with the mean of their scores analysed; perception was assessed with a stress-identification test marked in Blackboard. Analyses included paired-samples t-tests, Wilcoxon signed-rank tests, confidence intervals, paired effect sizes, inter-rater agreement statistics, and exploratory correlations. Mean production scores rose from 11.00 to 14.44 and mean perception scores from 11.12 to 15.42 (each out of 20). Both differences were statistically significant ($p < .001$), with standardized paired changes (*dz*) of 2.32 and 1.74, respectively. Production scores were higher at posttest for 25 of 26 students and perception scores for 24. The two raters agreed well on pretest and posttest production totals (ICC = .78 and .87) and moderately on individual change (ICC = .52), and the production gain was significant under each rater's scores separately. Without a comparison group, these results document improvement in recorded scores during the course; they do not isolate the effect of instruction. They nevertheless support assessing production and perception separately in university phonology courses for Arab EFL learners.

KEYWORDS

English stress; phonology instruction; production; perception; classroom assessment; Arab EFL learners

ARTICLE INFORMATION

ACCEPTED: 22 September 2026 **PUBLISHED:** 22 October 2026 **DOI:** 10.32996/ijllt.2026.9.10.24

1 Introduction

Knowing an English word involves more than knowing its meaning and the sounds from which it is formed. It also involves knowing which syllable carries prominence and recognizing that pattern when the word is heard. A learner may know a word in writing yet hesitate when encountering it in speech, or pronounce its individual sounds recognizably while placing stress on an unexpected syllable. Word stress therefore connects lexical knowledge with the practical demands of speaking and listening. Levis (2018) explains that stress patterns contribute to spoken-word recognition and that misplaced stress can direct listeners toward unintended lexical candidates. Attention to stress is consequently relevant to intelligible communication, not merely to conformity with a pronunciation model.

English lexical stress concerns the relative prominence of syllables within a word. Its realization involves a combination of phonetic cues, including duration, intensity, pitch, and vowel quality; these cues should not be reduced to a requirement that the stressed syllable must always be the loudest or highest-pitched. At the sentence level, prominence operates within a larger prosodic structure and interacts with information focus and intonation. Word stress and sentence prominence are thus connected but

AL-KINDI

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

distinguishable aspects of pronunciation. Roach's (2009) treatment of stress, weak forms, connected speech, and intonation provides the instructional foundation for approaching these relationships in the phonology classroom.

For learners, identifying where stress belongs and realizing it in speech present different demands. Perception requires attention to prominence in an incoming signal; production requires coordinating the relevant contrasts while speaking. Explicit knowledge may support both processes, but knowledge of a stress rule does not ensure that the learner can recognize or produce its realization consistently. This distinction gives pronunciation instruction a dual responsibility: to develop informed awareness of stress patterns and to provide practice in using that awareness during listening and speaking.

Recent intervention research offers evidence that these abilities can respond to instruction. Couper (2022) reported improvements in word-stress perception following conceptual instruction, including maintenance at delayed testing. Chen et al. (2025) found benefits from both rule-based and acoustic-perceptual instruction, with differences across outcome types and assessment occasions. However, improvement need not occur equally in both modalities. Deng et al. (2024), for example, reported production gains following song-based instruction without corresponding evidence of improved perception. Together, these findings support evaluating perception and production separately rather than combining them into a single undifferentiated pronunciation outcome.

Research involving Arab learners also provides a basis for focused classroom investigation. Yahy (2022) reported benefits from explicit stress-rule instruction among university EFL students in Iraq. Research focused specifically on Saudi female university students and vocational trainees has also identified lexical stress and rhythm among persistent pronunciation challenges (Balla & Elmahdi, 2024). Alzi'abi (2025) demonstrated improvement in the production of stress in suffixed words following targeted instruction among Arab EFL learners. These studies establish that the topic has an existing regional evidence base. They also raise a practical question for university teaching: how does stress performance develop when explanation, audio practice, production activities, and feedback are combined within a regular phonology course?

The present study examines an intact class of 26 Level 6 English majors enrolled in Phonology 321 at Taibah University in Madinah, Saudi Arabia. Their previous study of linguistics, literature, and language skills provided an academic foundation for phonological learning. The instructor described their general proficiency as approximately upper-intermediate; this description was not established through a standardized proficiency test. Over approximately eight weeks, two weekly 90-minute meetings combined textbook-based explanation with audio materials, perception and production practice, and corrective feedback. The course used Roach (2009). The approximately 24 scheduled contact hours covered the course as a whole, rather than stress instruction alone.

The research problem concerns whether this ordinary instructional setting was accompanied by measurable changes in students' stress performance, and whether those changes followed similar patterns in speaking and listening. A total production score samples lexical stress and sentence-level prominence across elicited tasks; a stress-identification score samples a narrower receptive ability. Neither score is a comprehensive measure of communicative competence. In this paper, productive and receptive awareness denotes performance on these administered tasks, not a separate measure of metalinguistic knowledge.

The study addresses a contextual gap rather than an absence of research on stress instruction. Earlier interventions differ in learner populations, teaching conditions, assessment formats, and follow-up periods. Their findings motivate the present investigation but cannot determine the outcome of a particular Saudi university course. Evaluating both modalities in the same cohort makes it possible to examine average change, variation among learners, and the association between outcomes. The contribution is therefore a classroom evaluation of assessed development, rather than a controlled estimate of the causal effect of a teaching method.

1.1 Research questions

Four questions guide the analysis:

RQ1. How do recorded English stress production scores change between the pretest and posttest?

RQ2. How do recorded stress-perception scores change over the same instructional period?

RQ3. How do the magnitude and variability of recorded score changes compare descriptively across the two assessments?

RQ4. How are production and perception scores associated at each assessment point, and how are their changes associated?

The two primary statistical tests evaluate the null hypothesis of zero mean within-student change separately for production and perception. Both are two-sided, with Holm adjustment across the two comparisons. Questions about cross-assessment differences and correlations are exploratory. A difference in recorded marks is distinguished throughout from a difference in underlying ability: equal score maxima do not establish equivalent instruments or equal measurement units.

2 Previous studies

2.1 Scope of the review

This targeted narrative review brings together recent studies of English stress instruction, with related work on assessment, feedback, and learner difficulties. It emphasizes publications from 2021 to 2026 and distinguishes intervention studies from diagnostic and methodological evidence. Research on German lexical stress is included as adjacent evidence about feedback, rather than direct evidence about English. The synthesis is organized by instructional approach and assessment purpose; it is not a systematic review or a quantitative aggregation of effects.

The main questions are whether instruction is associated with better stress performance, whether perception and production respond similarly, and whether improvements extend beyond practiced material or persist at delayed assessment. Studies with comparison conditions address different questions from one-group classroom evaluations. Both contribute to pedagogical understanding, but their raw scores and effect estimates cannot be treated as interchangeable.

2.2 Evidence from Arab EFL contexts

Abdulrahman and Ramamoorthy (2021) examined erroneous English stress patterns among 120 Yemeni EFL learners using recordings, auditory analysis, and Praat. Four-syllable words were especially problematic, and the reported patterns included stress on initial syllables and syllables containing long vowels or diphthongs. Although the paper uses experimental terminology, its contribution here is diagnostic: it identifies difficulties rather than testing a teaching intervention through a treatment–control comparison. It provides a regional rationale for focused instruction, while also showing why teachers should assess learners' actual stress patterns instead of assuming that general English proficiency guarantees accurate stress.

Yahy (2022) investigated explicit teaching of English stress rules among university EFL students in Iraq using experimental and control groups with pretest and posttest assessment. The study reported an advantage for explicit instruction over listening without explicit rule teaching. Its relevance lies in combining an Arab university context with a comparison between instructional conditions. The present course similarly incorporated explicit explanation, although its single-cohort design evaluates change over time rather than the relative effectiveness of explanation and listening.

Baagbah and Ganapathy (2025) examined Praat-supported English lexical stress production among 80 learners from Saudi Arabia and Yemen. The intervention combined visual acoustic feedback with self-directed learning, and pretest–posttest assessment examined duration, intensity, and fundamental frequency. The authors reported improvements in the experimental condition, particularly in duration, while interviews highlighted autonomy alongside the continuing importance of teacher guidance. This study offers regional evidence for making stress cues observable during practice. Nevertheless, acoustic improvement on particular dimensions should not be treated as interchangeable with higher overall classroom pronunciation marks or improved listening comprehension.

Alzi'abi (2025) investigated instruction targeting stress in suffixed words among Arab EFL learners. The production experiment involved 65 experimental and 60 control participants and a three-week instructional programme. Production improved following targeted teaching. A particularly important methodological distinction is that perception was assessed once, whereas production was assessed before and after instruction. Consequently, the study supports conclusions about production change but does not establish a pretest–posttest perception gain. Its contrast between relatively successful perception and weaker production also cautions against treating knowledge of stress placement as evidence that learners can reliably realize it in speech.

Alzi'abi (2026) extended this line of inquiry through a study of 180 Jordanian Grade 11 learners. Contrastive phonological analysis was compared with auditory discrimination training across different lexical stress patterns. Both approaches improved perception and production, with contrastive analysis showing an advantage for ambiguous patterns and delayed assessment indicating moderate retention. This study is valuable because it compares two plausible instructional approaches and examines persistence beyond immediate testing. However, the learners were secondary-school students; the findings cannot be transferred unchanged to university English majors with prior linguistics coursework.

A complementary Saudi university perspective comes from Elmahdi and AbdAlgane (2024), who surveyed 33 EFL professors across Saudi universities about pronunciation instruction. Their findings point to varied instructional priorities and a need for better-tailored pronunciation teaching. Because the study reports instructors' perspectives rather than a stress-specific student intervention, it contextualizes the present course but does not provide a direct benchmark for its score changes.

Taken together, these studies establish a regional basis for investigating stress instruction and show that Arab EFL learners can benefit from targeted practice. They also demonstrate the importance of distinguishing receptive recognition from spoken realization. The present study extends this line of classroom inquiry to an intact Saudi university cohort, with both outcomes assessed during a regular phonology course.

2.3 Explicit instruction and instructional sequence

Couper (2022) examined a conceptual approach to English word-stress perception among high-intermediate adult learners in New Zealand. The analysis included 18 participants: ten in the instructional group and eight in the comparison group. Four lessons over four weeks helped learners develop and articulate concepts of stress rather than simply imitate models. Perception improved in the instructional group and gains were maintained at an eight-week delayed assessment, while the comparison group showed little progress. The proficiency level and classroom setting make this study relevant to Taibah students, although the small, multilingual migrant sample differs from an intact Saudi university class.

Chen and Tian (2024) studied metaphonological awareness protocol training with 40 Chinese EFL learners at different proficiency levels. Pretest–posttest results indicated improvement in both higher- and lower-proficiency learners, with larger gains among the lower-proficiency group. The study also considered awareness and self-repair, connecting instruction with learners' ability to explain or monitor stress decisions. It supports including reflection alongside practice, but proficiency-group differences should not be mistaken for evidence from an untreated control group. For the present study, it suggests that prior English learning and initial stress performance may influence how much students gain.

Chen, Han, and Tian (2025) compared rule-based instruction, acoustic-perceptual instruction, and regular instruction with 90 Chinese ESL learners allocated across three groups of 30. Eight instructional sessions were evaluated through pretests, posttests, and delayed tests involving reading aloud, stress identification, and vowel reduction. Both targeted approaches produced benefits. Rule-based teaching had a slight short-term production advantage, whereas acoustic-perceptual instruction offered benefits for perception and longer-term performance. The study is especially useful because it treats instructional effectiveness as dependent on the skill and assessment occasion, rather than identifying one universally superior method.

Zhou et al. (2025) investigated instructional sequence with 80 Chinese adolescent learners over five weeks. One group received explicit instruction followed by implicit instruction, while the other experienced the reverse sequence. Pretest, immediate posttest, and delayed posttest measures included word-stress identification accuracy and response time. Both groups improved, but the implicit-to-explicit sequence showed more persistent benefits. Here, blended instruction refers to combining instructional approaches, not necessarily to combining online and classroom delivery. The study highlights that the order of explanation and practice may matter, although it does not establish an optimal sequence for every learner population.

The explicit-instruction studies suggest that stress teaching can combine conceptual explanation, auditory discrimination, production practice, and reflection. Their comparisons also show that instructional effects may depend on the outcome and assessment occasion. Because these components were integrated in Phonology 321, the present study evaluates their combined classroom context rather than estimating the independent contribution of each component.

2.4 Technology and independent practice

Maspufah (2022) explored YouTube videos for teaching English word-stress pronunciation to 17 students. The classroom action research used tests, interviews, observation, and field notes and reported improvement during the intervention. Its practical contribution is to show how accessible video material can be incorporated into classroom pronunciation work and examined using several sources of evidence. Its design, however, differs from a controlled experiment: changes during classroom action research may reflect the combined effects of instruction, repeated practice, and revisions to teaching. It therefore offers supportive classroom evidence rather than an isolated estimate of the effect of YouTube.

Sain and Cobar (2023) investigated the text-to-speech feature of Google Translate with 22 Indonesian Grade 11 students. Their pre-experimental, one-group study involved five meetings, including assessment and treatment. Mean pronunciation scores increased from 32.04 to 60, and the authors reported a statistically significant paired comparison. The design closely resembles classroom investigations that assess one cohort before and after instruction. Its limitation is equally relevant: without a comparison group, the improvement cannot be attributed exclusively to the technological tool. Moreover, its raw scores should not be compared directly with marks out of 20 because the assessment scales differ.

Sardegna and Jarosz (2023) examined autonomous technology-supported stress learning with six motivated Polish secondary-school students. Over four weeks, learners used orthographic stress rules, worksheets, and YouGlish, with read-aloud pretests and posttests and activity logs. Accuracy improved for the two rule categories investigated. The combination of rule explanation and examples from authentic speech offers a useful model for independent practice. Nevertheless, the small, selected sample and absence of a control group restrict generalization. The study is a conference-proceedings contribution and should be treated as promising exploratory evidence rather than a large-scale efficacy trial.

Hirschi and Kang (2024) compared corpus-based pronunciation learning with teacher-created materials among 65 university learners of English in North America. Forty learners used corpus materials and 25 formed the comparison condition across four online lessons. Listening and speaking assessments addressed lexical stress and prominence. Both groups improved lexical stress production; the corpus condition additionally showed benefits in prominence production and recognition. The results therefore

do not justify claiming that corpus materials were superior for every outcome. The study is useful for distinguishing lexical stress from phrase-level prominence and for showing why production and perception should be analysed separately.

Ivanova (2024) reviewed technology for training English word-stress perception and production and discussed a proposed training tool incorporating high-variability phonetic training. The paper is relevant to the design of pronunciation resources because it emphasizes stress-specific learning needs rather than assuming that any pronunciation application addresses them. It is not an intervention trial demonstrating the effectiveness of the proposed tool. Accordingly, it provides conceptual and technological background, while claims about learning gains must rest on empirical studies. This distinction is particularly important when discussing newer tools whose pedagogical appeal may exceed the available outcome evidence.

2.5 Multimodal support

Chayluxsamee and Lynch (2022) compared multisensory instruction with a listen-and-repeat approach in two intact Grade 5 classes in Bangkok. The eight-week programme incorporated short word-stress activities twice weekly, with pretest and posttest assessment. The multisensory condition produced greater improvement than the comparison approach. The duration resembles the present study, but the amount of stress-specific teaching and the learners' age differ substantially. Its main contribution is evidence that adding structured sensory support can benefit stress learning beyond repetition alone; it should not be used to predict the size of gains among upper-intermediate university students.

Nakplad et al. (2023) used an animated film and syllable-stress exercises with 30 Thai primary-school students. Oral pretest–posttest performance improved significantly following the activities. This study illustrates how audiovisual materials can provide a context for stress practice, especially with younger learners. However, the intervention combined the film with exercises, and a single-cohort comparison cannot separate the contribution of the film from that of guided practice or repeated assessment. For a university phonology course, the relevant principle is the integration of listening examples with purposeful stress activities, rather than the assumption that entertainment material itself produces learning.

Deng et al. (2024) investigated English songs with 39 students in a rural Chinese junior high school over two weeks. The study assessed perception and production of disyllabic word stress and included acoustic analysis. Production improved, particularly the duration contrast associated with stress, whereas perception did not show improvement. This finding is important because it challenges the assumption that listening-based activities necessarily produce parallel gains in listening and speaking. It supports treating the two outcomes as related but distinct and examining the actual assessment results rather than predicting improvement from the apparent modality of the teaching material.

Helal et al. (2024) compared conventional word-based captions with phonetic transcriptions in video-supported stress instruction. The study used pretest–posttest assessment with two EFL groups and did not find a statistically significant posttest difference between caption conditions. This result should be represented accurately: it does not establish superiority of phonetic transcription, and a nonsignificant difference is not proof that the two approaches are equivalent. The study is useful for tempering strong claims about specialized visual support and for demonstrating that an instructional feature can be plausible without showing a clear comparative advantage in a particular sample.

Kim and Archibald (2024) examined lexical stress diacritics with 30 Korean speakers of English. Learners read sentences before instruction, then received explicit guidance and read marked sentences, followed by an unmarked delayed assessment two days later. Native-speaker ratings indicated improvements in comprehensibility and accentedness. The study links stress support with listener judgments rather than stress-location accuracy alone. Nevertheless, the immediate posttest included visual cues absent from the pretest, so part of the performance difference may reflect assistance during testing. The brief delayed assessment is useful but should not be described as evidence of long-term retention.

Yang et al. (2026) compared gestures signalling duration, intensity, or pitch during English lexical stress training with 79 Chinese adolescents. Pretest–posttest measures included perception and acoustic production across stress and intonation patterns. Duration-signalling gestures showed the strongest generalization to untrained material, while some combinations of stress and intonation remained more difficult. The study demonstrates why multimodal support should be evaluated in relation to the particular cue it emphasizes. It also provides a useful assessment model: performance on unfamiliar items can reveal whether learning extends beyond practiced examples, although the findings remain specific to the population and conditions investigated.

2.6 Instruction within university and professional courses

Thongsongsee and Watanapokakul (2023) analysed the production of stress in English medical terminology by 75 first-year Thai medical students. The assessment involved 40 polysyllabic terms, supplemented by interviews with nine participants. The study identifies a need for attention to stress in specialized vocabulary, but it does not test an instructional intervention. Its relevance is therefore diagnostic and curricular. It suggests that familiarity with an academic field and access to technical vocabulary should not be assumed to ensure accurate pronunciation, while leaving the effectiveness of any proposed teaching response to be established through intervention research.

Kanoksilapatham and Pariwatnanont (2024) integrated brief English word-stress instruction into a Thai university business course involving 52 students. Instruction occupied five to eight minutes in each of eight lessons, and assessment used 40 business words before and near the end of the course. The authors reported improved stress placement and positive student responses. The study closely supports the practicality of embedding pronunciation in ordinary coursework. Its results nevertheless concern a combined classroom experience; without a separate comparison condition, they do not determine how much improvement would have occurred through the course and repeated exposure alone.

Kanoksilapatham (2025) investigated stress-focused teacher development with 49 university instructors working across seven disciplines in English-medium instruction. Training addressed 80 words associated with classroom functions, and pretest–posttest results showed increased stress knowledge. The investigation also considered production and participants' evaluations of the training. It broadens the literature beyond student samples and demonstrates the relevance of stress to professional classroom speech. However, positive perceptions of a training programme are attitudinal findings, not evidence of auditory stress-perception gains. Teacher-development outcomes should therefore be distinguished from the listening outcomes assessed in the present student study.

Kanoksilapatham and Takrudkaew (2026) investigated explicit lexical stress instruction alongside TOEIC preparation with 73 fourth-year Thai mechanical-engineering students over 15 weeks. Assessment included a 50-item stress test and mock TOEIC listening tests before and after instruction. Both outcomes improved. This study connects focused pronunciation work with a broader listening measure, but the combined intervention prevents attributing TOEIC improvement exclusively to stress teaching. It is therefore relevant to the possibility of wider listening benefits while also illustrating the need to distinguish stress identification from general listening comprehension and to describe all instructional components.

2.7 Feedback retention and assessment

Sippel and Martin (2023) provide adjacent evidence from German lexical stress learning. Their study involved 82 first-year learners across 11 course sections and compared teacher feedback, providing peer feedback, receiving peer feedback, and a control condition. Pretest, immediate posttest, and delayed posttest results indicated improvement for teacher feedback and peer-feedback providers, but not for receivers or the control condition. Because German was the target language, this is not direct evidence about English stress instruction. Its contribution is a testable pedagogical possibility: actively evaluating pronunciation may support learning differently from simply receiving corrections.

Hosaka and Saito (2026) investigated instructed suprasegmental learning with 115 Japanese learners of English. Six sessions compared perceptual training with explicit instruction and feedback against an explicit-instruction-only condition. Assessment included trained and untrained words. Improvement was clearer for trained material, with limited generalization, and non-pitch auditory processing was associated with learning gains whereas pitch processing was not. The study strengthens the case for examining unfamiliar items and individual differences. It also cautions against treating success on repeated test words as sufficient evidence that learners have acquired a broadly transferable stress skill.

Duris et al. (2024) examined how assessment method influences English stress identification. Three expert listeners evaluated 1,000 words produced by ten Chinese speakers. Agreement was higher for analytic identification of primary stress than for a global judgment of stress correctness. The distinction matters for classroom assessment: identifying the syllable that a speaker makes prominent is a different judgment from deciding whether the whole word is pronounced correctly. The present production totals represent instructor judgments and do not include an independently estimated inter-rater reliability coefficient.

2.8 Synthesis and implications

Across the reviewed studies, targeted stress teaching is frequently associated with improved assessed performance. The findings are most informative when the instructional contrast and the measured outcome are specified. Explicit rules, perceptual practice, visual support, and feedback may contribute in different ways, and success in one task does not ensure improvement across all tasks. Comparison groups and delayed testing strengthen conclusions about instructional contributions and retention; smaller classroom studies establish feasibility and describe local development.

A central implication concerns the relationship between perception and production. Listening to a model, locating primary stress, producing a stress contrast, and speaking comprehensibly draw on related skills but impose different demands. Separate assessment is therefore necessary. The present analysis retains two outcomes and examines their association without treating a correlation as evidence that one skill causes development in the other.

Transfer and retention require similarly careful distinctions. Familiar test words may reflect learning alongside practice and memory. New words test additional material but can introduce changes in difficulty. Sentence and passage reading add demands beyond isolated lexical stress. In the present study, the immediate post-instruction scores document performance on the administered forms; they do not establish durable retention or generalization to spontaneous conversation.

Phonology 321 provides a relevant setting for studying these issues because students encounter explicit phonological concepts alongside listening, spoken practice, and feedback. The class had prior academic experience in English, yet its baseline scores left substantial room for higher assessed performance. Following the same students across the teaching period links course evaluation to individual outcomes as well as cohort means.

The analytical response is to report raw score changes, their confidence intervals, and paired standardized changes together. Mean differences show the magnitude of change in recorded marks; individual trajectories show its distribution; correlations examine whether the two assessments rank students similarly. This combination supports a more informative evaluation than significance tests alone, while preserving the distinction between documented score change and a causal account of learning.

3 Method

3.1 Design and participants

The study used a quantitative single-group pretest–posttest design within the regular Phonology 321 course. The participants were the entire class of 26 Level 6 English majors at Taibah University. All students had complete production scores from both raters and complete perception totals at both assessment points, and all were included in the analysis. This was a census of one instructional cohort, not a random sample of the university population. There was no control group or random allocation.

Before taking Phonology 321, students had studied courses in linguistics, literature, and English language skills. The instructor described their general English proficiency as approximately upper-intermediate. This description reflects an instructional judgement rather than a proficiency classification established by a standardized test administered for the study. The participants' previous coursework provided a foundation for the course, but their initial command of stress was assessed through the pretest rather than assumed from their academic level.

3.2 Instructional procedures

The instructional period lasted approximately eight weeks. The class met twice weekly for 90 minutes per meeting, corresponding to approximately three hours per week and 24 scheduled contact hours across the study period. These hours represent the course schedule; they should not be interpreted as 24 hours devoted exclusively to stress instruction.

The principal textbook was the fourth edition of Peter Roach's English Phonetics and Phonology. Teaching incorporated extensive use of audio materials and practice in stress production and perception. Students listened to spoken examples, practised identifying stress, and undertook production activities. The instructional emphasis combined explanation of stress with repeated listening and opportunities to practise its realization in speech.

The instructor provided extensive corrective feedback throughout the teaching period. Feedback addressed students' performance during practice and supported further attention to the stress patterns being taught. The instructional condition is therefore described as regular course-based teaching with substantial audio practice and teacher feedback. It was not a separate AI-delivered intervention: the reported use of AI concerned preliminary assessment of production recordings rather than delivery of the course.

Table 1 Study procedures

Stage	Timing	Procedure
Pretest	Beginning of the semester	Assessment of baseline production and perception performance
Instruction	Approximately eight weeks	Two weekly 90-minute meetings using the textbook, audio materials, practice, and corrective feedback
Posttest	After the instructional period	Reassessment of production and perception performance
Scoring and analysis	Following assessment	Production scores from two independent raters (averaged) and Blackboard-marked perception scores paired within students

Note. The scheduled contact time includes all course content, not exclusively stress instruction.

3.3 Production assessment

The production pretest sampled individual words, sentence reading, and passage reading. Its ten word items carried ten marks, five sentences carried five marks, and the passage carried five marks. Examples of the word material were table, development, participation, alone, academic, explanation, important, economic, tomorrow, and responsibility. Sentence tasks elicited prominence in connected sequences, including "She works for me in the office." The pretest maximum was 20 marks.

The production posttest sampled word reading, sentence reading, and passage reading, including retained and additional material. Additional words included indigent, prominent, miserable, apart, phenomenology, apology, deliberation, environmental,

generosity, and demonstrate. Some sentence material was repeated or adapted, while the longer passage included academic vocabulary. Instructions requested natural pace, no artificial pauses between syllables, and repetition of an item after a mistake. The production posttest was scored out of 20. The additional words were drawn from material studied in the course.

Table 2 Production assessment tasks and maximum scores

Assessment	Word reading	Sentences and passage	Maximum score
Pretest	10 words	5 sentences and a passage	20
Posttest	Retained and additional words	Sentence and passage reading	20

Note. Both production assessments were scored out of 20.

The analytical outcome was the production score out of 20 at each assessment occasion, defined as the mean of the two raters' scores (Section 3.5); it can therefore take half-mark values. The analysis used these scores directly. Because the posttest included additional words drawn from course material, results for repeated and additional material cannot be separated from the overall totals.

The tasks targeted lexical stress, primary and secondary prominence where appropriate, and sentence rhythm. The sentence instructions used final-content-word prominence as the elicitation target. This is a task-specific convention for the intended reading, not a claim that English nuclear prominence invariably falls on the last lexical word: discourse focus and intonational phrasing can alter its placement (Roach, 2009). No acoustic measurements of pitch, intensity, duration, or vowel reduction were entered into the analysis.

3.4 Perception assessment

The receptive outcome was stress identification, referred to as listening in the grade sheet. The posttest form contained 20 isolated-word items in two sets of ten. Students heard one word at a time and selected the syllable position perceived as most prominent. The first set included important, academic, economic, development, certificate, Japanese, generosity, participation, environmental, and explanation. The additional set included Portuguese, experiment, deliberation, apologetic, generous, photographic, hallucinate, demonstrate, mountaineer, and permit as a verb.

The written response form displayed five syllable-position options. In its first set, the fourth and fifth position labels appeared in reverse order relative to the second set. Because responses were scored in Blackboard, this paper does not assume that the paper layout was reproduced in the electronic answer key. The analytical record contains finalized totals rather than the Blackboard item configuration. Both the perception pretest and posttest were scored out of 20; the analysis does not establish item-level equivalence between forms. This assessment represents prompted stress identification, not general listening comprehension or an independent measure of conscious phonological knowledge. One perception posttest total was corrected after verification against the Blackboard record before analysis.

3.5 Scoring and measurement

The production pretest, production posttest, listening pretest, and listening posttest were each scored out of 20. Production recordings were scored independently by two raters: the course instructor (first author) and the second author. The second author did not have access to the first author's scores. The first author used Gemini and ChatGPT to obtain preliminary assessments of the recordings, after which he carried out the final correction; he reported only minor differences between the preliminary AI assessments and his own judgements, and whenever a discrepancy occurred his judgement determined his score. The production score analysed for each student and occasion is the mean of the two raters' scores.

Agreement between the raters is summarized in Table A1 (Appendix A). Agreement was high for pretest and posttest production totals (ICC(2,1) = .78 and .87; Pearson r = .79 and .86) and moderate for change scores (ICC(2,1) = .52; r = .53). The raters differed by no more than one mark for 24 of 26 students at pretest and 25 of 26 at posttest, but identical totals were uncommon (1 and 2 students, respectively), and their change scores were within one mark for only 10 students. Neither rater's scores differed systematically from the other's (bias tests p = .188, .574, and .204). The AI systems were not treated as independent raters or as a validated reference standard. Because both raters were involved in the study and knew its purpose, and no external blinded rater was used, production scores remain subject to rater effects. Using the mean of two raters reduces dependence on any single judgement, and results for each rater separately are reported in Table A2 (Appendix A).

Listening responses were marked directly by Blackboard. AI assistance was not used for this component. Production and listening therefore differed in their marking procedures: production required two raters' evaluation of spoken performance, whereas listening marks were generated by Blackboard from students' test responses. The second author's listening records are identical to the Blackboard totals, because no separate human marking of listening was undertaken.

The retained scores (each rater's production totals and the Blackboard perception totals at both occasions) support analysis of score distributions, within-student change, and agreement between raters on totals. They do not support internal-consistency

estimates, task-specific effects, or item-level agreement. These measurement limits apply to the interpretation of the results; they are not replaced by claims of reliability based on automated marking.

3.6 Statistical analysis

Records were matched within students. All 26 records contained production scores from both raters and perception scores at both occasions (156 observed totals, or 104 analytical totals after averaging the raters' production scores), with no missing values. All recorded values fell between 0 and 20. Descriptive statistics included the mean, sample standard deviation, median, and observed range. Change was defined as posttest minus pretest, computed from the analytical scores. The number of students with higher, unchanged, and lower scores was calculated separately for each outcome.

Two-sided paired-samples t-tests evaluated mean changes in recorded scores. Holm adjustment controlled multiplicity across these two primary tests at an alpha level of .05. Mean changes are accompanied by paired t-based 95% confidence intervals. Cohen's d_z is the mean within-student change divided by the sample standard deviation of those changes. It measures standardized change in this cohort and is not an effect size relative to an untreated comparison group. For comparison with studies that standardize by pooled variability, d_{av} (mean change divided by the mean of the pretest and posttest standard deviations) is also reported. The primary tests used the mean of the two raters' production scores; the same tests were repeated for each rater separately as a sensitivity analysis.

Shapiro–Wilk tests assessed the distributions of the change scores. Wilcoxon signed-rank tests were conducted as complementary rank-based checks, using a two-sided normal approximation with adjustment for ties and excluding zero differences from ranking. These analyses were used to evaluate the consistency of the findings across approaches, while recognizing that a signed-rank test does not directly test a mean change and that a median-shift interpretation requires symmetry.

Exploratory Pearson correlations examined the two outcomes at pretest, at posttest, and as change scores. Their 95% confidence intervals use Fisher's transformation. An additional paired comparison describes the difference between changes on the two recorded 20-mark scales. Because the assessments are not equated, this comparison does not test whether underlying productive and receptive abilities improved equally. Exploratory p values are reported without multiplicity adjustment. Figures display individual trajectories, mean changes with confidence intervals, and score-pair scatter plots; overlapping pairs are represented by marker area rather than jittered coordinates.

Agreement between the two raters' production scores was evaluated at pretest, at posttest, and for change scores. Statistics were two-way random-effects, absolute-agreement, single-rater intraclass correlation coefficients (ICC(2,1)) with F-based 95% confidence intervals; Pearson correlations with Fisher confidence intervals; quadratic-weighted kappa; the mean and standard deviation of rater differences with 95% limits of agreement; the mean absolute difference; and counts of exact agreement and agreement within one mark. A one-sample t-test of the differences assessed systematic bias between raters. These analyses are descriptive and unadjusted.

3.7 Ethics and use of assessment records

The assessments formed part of routine course assessment. Students were not informed at the time that their scores or recordings would be used for research, and research-specific informed consent was not obtained. No institutional ethical approval or approved waiver is documented. Course participation is not treated as consent to research use. Production recordings were also subjected to preliminary AI-assisted assessment before the instructor finalized the marks. These procedures are disclosed as conducted.

The manuscript reports aggregate results and unlabeled graphical observations, with no student names or identification numbers. Identifiable records and recordings are not reproduced. Removing identifiers from the report does not retrospectively establish consent or ethical approval for the underlying research use.

4 Results

4.1 Score distributions

Table 3 Descriptive statistics for recorded scores

Assessment	N	Mean	SD	Median	Range
Production pretest	26	11.00	1.67	11.50	8.5–13.5
Production posttest	26	14.44	1.89	14.50	9.5–17.5
Perception pretest	26	11.12	2.12	11.50	5–15
Perception posttest	26	15.42	2.93	16.50	8–20

Note. N = 26. All four tests were scored out of 20. Production scores are the mean of the two raters' scores. SD denotes the sample standard deviation.

Mean production scores (the mean of the two raters' scores) increased from 11.00 (SD = 1.67) to 14.44 (SD = 1.89). The median increased from 11.50 to 14.50, and the range shifted from 8.5–13.5 to 9.5–17.5. Perception scores increased from a mean of 11.12 (SD = 2.12) to 15.42 (SD = 2.93), with medians of 11.50 and 16.50. The perception posttest range was 8–20. Thus, higher average scores were accompanied by increased dispersion in both outcomes, more markedly in perception.

On the recorded 20-mark scales, the production means correspond to 55.00% and 72.21% of the stated maximum, and the perception means to 55.58% and 77.12%. These percentages describe the grade-sheet totals. They are not percentages of mastery or evidence that the underlying assessments are equally difficult.

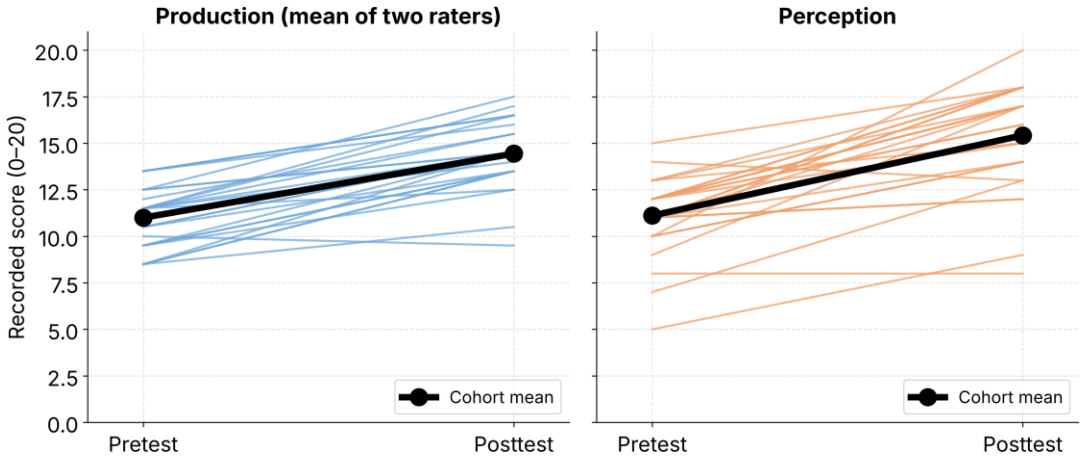
4.2 Individual changes

Table 4 Direction of changes in recorded scores

Outcome	Higher score	Unchanged	Lower score
Production	25 (96.2%)	0 (0%)	1 (3.8%)
Perception	24 (92.3%)	1 (3.8%)	1 (3.8%)

Note. Percentages use all 26 students. Higher or lower scores refer to the recorded totals, not independently validated changes in ability.

Production scores (the mean of the two raters' scores) were higher at posttest for 25 of the 26 students; one student's score was 0.5 marks lower, because the first rater's total fell by one mark while the second rater's was unchanged. In perception, 24 students had higher scores, one was unchanged, and one had a lower score (a fall of one mark). The cohort means therefore summarize different individual patterns: production changes were positive for almost all students and varied little, whereas perception gains varied more widely across students.



Paired trajectories for 26 students showing recorded production (mean of two raters) and perception scores at pretest and posttest.

Figure 1 Individual score trajectories and cohort means

Note. Thin lines connect paired observations; black lines show cohort means. Overlapping trajectories may conceal individual lines. N = 26. Production scores are the mean of the two raters' scores. Both panels display test scores out of 20.

4.3 Magnitude and uncertainty of score changes

Table 5 Paired comparisons of recorded scores

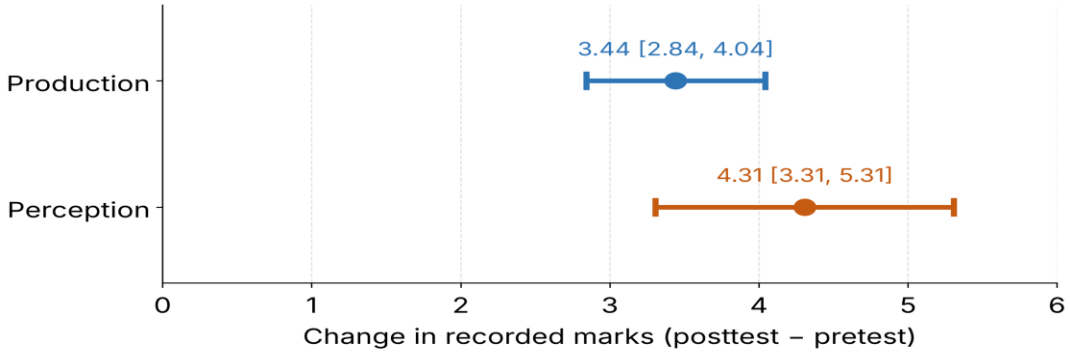
Outcome	Mean gain	SD gain	95% CI	t(25)	p	d_z
Production	3.44	1.49	[2.84, 4.04]	11.82	< .001	2.32
Perception	4.31	2.48	[3.31, 5.31]	8.86	< .001	1.74

Note. Change = posttest minus pretest. CI = confidence interval. Both two-sided tests remain significant after Holm adjustment. d_z = mean change / SD of change. Production scores are the mean of the two raters' scores. Production change is interpreted at the level of recorded scores.

The mean production-score change was 3.44 marks (SD = 1.49), with a 95% confidence interval of [2.84, 4.04]. The paired comparison was significant, t(25) = 11.82, p < .001, d_z = 2.32. The mean perception-score change was 4.31 marks (SD = 2.48), with

a 95% confidence interval of [3.31, 5.31], t(25) = 8.86, p < .001, and d_z = 1.74. Holm-adjusted p values were also below .001 for both outcomes. Standardizing by the mean of the pretest and posttest standard deviations instead gave d_{av} = 1.93 for production and 1.71 for perception. The production result was similar for each rater's scores analysed separately (first rater: mean change 3.65, 95% CI [2.98, 4.33]; second rater: 3.23, 95% CI [2.53, 3.93]; both p < .001; Table A2).

The larger standardized production change reflects its smaller change-score standard deviation, not a larger mean increase. Its observed mean (3.44 marks) was below the perception mean (4.31 marks). This distinction is particularly relevant for paired effect sizes: a highly consistent set of changes can yield a large d_z even when the absolute increase is comparable with that of another assessment. Neither value isolates an instructional effect from changes in test form, scoring, or other semester experiences.



Mean recorded production and perception score changes with paired 95 percent confidence intervals.

Figure 2 Mean changes in recorded scores with 95 percent confidence intervals

Note. Points represent mean changes and horizontal bars represent paired t-based confidence intervals. Values are changes in recorded marks. Interval widths reflect sampling variability, not uncertainty about score calibration or test equivalence.

4.4 Distributional checks

Table 6 Change-score normality and signed-rank tests

Outcome	Shapiro–Wilk W	p	Wilcoxon T	p
Production gain	.954	.287	1.00	< .001
Perception gain	.977	.811	2.00	< .001

Note. Wilcoxon tests use the two-sided normal approximation with tie adjustment and no continuity correction. The zero perception difference is excluded from ranking. T is the smaller signed-rank sum.

Shapiro–Wilk tests did not indicate departures from normality for production changes, W = .954, p = .287, or for perception changes, W = .977, p = .811. The Wilcoxon tests provide a rank-based check, and they rejected their null hypotheses for both outcomes: production T = 1.00, p < .001; perception T = 2.00, p < .001. The positive signed-rank sums predominated in each case. These analyses complement the paired mean comparisons but do not resolve measurement differences between assessments. A signed-rank test is not a direct test of mean change, and a median-shift interpretation depends on symmetry of the difference distribution.

4.5 Exploratory comparisons and associations

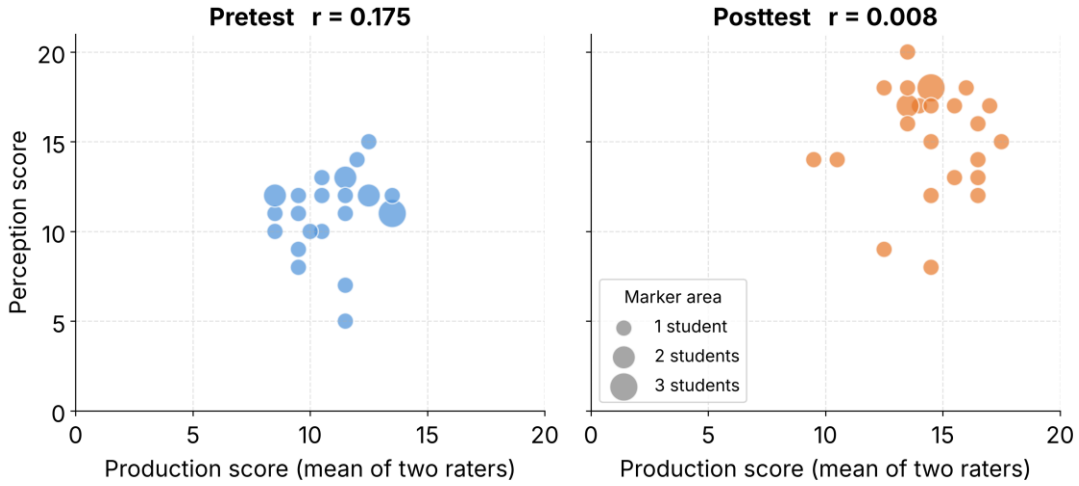
Subtracting each student's production change from their perception change yielded a mean difference of 0.87 recorded marks, with a 95% confidence interval of [−0.36, 2.09]. The paired comparison was nonsignificant, t(25) = 1.46, p = .157. This describes the numerical relationship between the two recorded scales. It does not establish equivalent learning across modalities, either statistically or psychometrically.

Table 7 Exploratory correlations between recorded outcomes

Relationship	r	95% CI	p
Pretest production and perception	.175	[−.228, .526]	.394
Posttest production and perception	.008	[−.380, .394]	.968
Production gain and perception gain	−.109	[−.476, .291]	.596

Note. N = 26 for each Pearson correlation. Confidence intervals use Fisher's transformation. Two-sided p values are unadjusted because these analyses are exploratory.

The pretest association was small and positive, r = .175, p = .394; the posttest association was close to zero, r = .008, p = .968. The correlation between changes was also small and negative, r = −.109, p = .596. The corresponding confidence intervals were wide and included both negative and positive values. These findings provide no clear evidence of a linear association in this cohort; they do not establish independence of the underlying abilities.



Scatter plots of recorded production and perception scores at pretest and posttest; marker size represents coincident observations.

Figure 3 Production and perception scores at each assessment

Note. Marker area is proportional to the number of students at the same coordinate. Axes show recorded scores out of 20. Coordinates are not jittered. The displayed correlations summarize linear association only.

4.6 Agreement between raters

Table A1 (Appendix A) summarizes agreement between the two raters' production scores. At pretest the raters' totals correlated at r = .79 (ICC(2,1) = .78, 95% CI [.57, .90]) and at posttest at r = .86 (ICC(2,1) = .87, 95% CI [.72, .94]). Mean differences between raters were small (−0.31 marks at pretest and 0.12 at posttest) and not significantly different from zero (p = .188 and p = .574). The raters' totals were within one mark for 24 of 26 students at pretest and 25 of 26 at posttest, although identical totals were rare (1 and 2 students), and the largest difference was 3 marks at pretest and 2 at posttest. Agreement on change scores was lower (r = .53; ICC(2,1) = .52, 95% CI [.18, .75]), with the raters' change scores within one mark for 10 of 26 students and 95% limits of agreement of −2.82 to 3.66 marks. Individual change scores based on a single rater should therefore be interpreted cautiously, which is one reason the analysis uses the mean of both raters.

Although the differences in average level were not statistically significant, the second rater's pretest scores were slightly higher and posttest scores slightly lower than the first rater's, so the second rater's mean gain (3.23 marks) was smaller than the first rater's (3.65 marks; Table A2). Both gains were statistically significant (t(25) = 9.54 and 11.14, p < .001), and the conclusion that production scores rose did not depend on which rater's scores were used.

5 Discussion

5.1 Interpreting higher scores after instruction

The main finding is a clear upward shift in recorded classroom scores following the instructional period. Production scores increased for 25 of 26 students, while perception totals increased for 24. The paired mean comparisons and complementary rank analyses were consistent in rejecting no change in the recorded outcomes. This is useful evidence for evaluating the course: students completed the period with higher assessed performance than at its beginning. It is not, by itself, an estimate of how much improvement instruction caused.

The direction of the findings is consistent with studies reporting benefits from explicit or targeted stress teaching (Alzi'abi, 2025; Chen et al., 2025; Couper, 2022; Yahy, 2022). Those interventions differ in design, learner characteristics, duration, and instruments. The meaningful comparison is therefore that stress-focused learning can be reflected in classroom assessment, rather than that one study produced a larger effect than another. The present study adds evidence from students already engaged in formal English and linguistic coursework.

The course combined explanation, auditory models, production practice, and feedback. Such a combination plausibly supports the connection between identifying prominence and realizing it in speech, but the design cannot distinguish its active

components. The instructional interpretation is strongest as an account of the complete classroom experience. It would be inaccurate to assign the observed gains specifically to the textbook, feedback, AI, or any single practice activity. AI assisted preliminary assessment; it was not the instructional treatment.

Both production assessments were scored out of 20, and the observed mean increase was 3.44 marks. This result indicates improved performance on the production assessments following instruction. Because the posttest included retained and additional material, the overall scores do not distinguish performance on familiar items from performance on new items.

5.2 Differences between productive and receptive outcomes

The mean recorded changes were similar (3.44 and 4.31 marks), but their distributions differed. Production changes had an SD of 1.49, compared with 2.48 for perception. The production d_z of 2.32 follows mathematically from this concentration of changes; it does not show that production learning was more effective than perception learning. A large paired effect is compatible with a modest range of individual changes when their mean is consistently above zero.

Deng et al. (2024) reported production improvement without corresponding perception improvement after song-based instruction. Hirschi and Kang (2024) found that results depended on whether lexical stress or prominence was assessed. These findings help place the present results in context: modalities and task types need not respond uniformly to an instructional programme. Here, the recorded perception mean increased, but one student scored lower. The distribution warrants attention to individual receptive performance even when the class average is favourable.

The weak correlations do not demonstrate a dissociation between perception and production. With 26 students, their intervals permit a range of population relationships. Measurement error, restriction of score range, different task demands, and different marking procedures may also affect the correlations. The defensible pedagogical implication is to assess both outcomes directly. Inferring a learner's receptive progress solely from a production score would not be justified by these data.

5.3 Pedagogical implications

For Phonology 321, the findings support retaining explicit attention to stress alongside purposeful listening and speaking practice. Explanation can give learners a vocabulary for discussing prominence, while listening examples and oral practice provide opportunities to apply it. Feedback is most informative when it identifies the location and realization of prominence rather than simply labeling a whole word correct or incorrect. The distinction between analytic and global ratings described by Duris et al. (2024) is relevant to designing such feedback.

Assessment can also inform differentiated support. A student whose production total increases but whose perception total remains unchanged may need additional stress-identification practice; a higher perception score with weaker production may motivate further spoken rehearsal. These are instructional responses to observed profiles, not causal conclusions about why a particular student changed. Class-level means are useful for course evaluation, but paired individual records provide the basis for identifying learners who require closer attention.

The task distinction between lexical stress and sentence prominence should remain explicit in teaching and reporting. Isolated words help focus attention on within-word prominence, whereas sentences and passages introduce phrasing, focus, and weak forms. Final-content-word prominence provides a controlled reading target in these assessments, but instruction should also explain how discourse context licenses other nuclear placements. Future course assessments can retain separate subscores for these levels, making the results more diagnostically useful.

5.4 Limitations and future research

The single-class design limits causal inference and generalization. Without an untreated or alternative-instruction comparison group, concurrent learning, repeated exposure, and familiarity with assessment cannot be separated from course-related development. The instructor's judgment of upper-intermediate proficiency supplies context but is not an independently measured covariate. The findings concern this cohort and these assessment occasions, rather than all Saudi university English learners.

Measurement limits include differences in test material, absence of component-level scores, and the absence of an external rater who was blind to the study; both production raters were involved in the study, and their agreement on individual change scores was only moderate. The perception analysis likewise lacks item-level responses and the archived Blackboard key needed to examine reliability and response-option coding. Automated scoring ensures consistent application of a key, but it does not establish that the key or instrument is valid. Statistical robustness checks address aspects of score distribution; they cannot compensate for these limitations.

The study did not include delayed testing, spontaneous interaction, or acoustic analysis. It therefore provides no direct estimate of retention, communicative intelligibility, or changes in particular phonetic cues. Future work can use parallel forms, retain repeated and untrained items as separate outcomes, and combine independent analytic ratings with acoustic measures. A comparison condition and delayed assessment would strengthen the connection between course activities, immediate performance, and sustained learning.

Research use of the routine assessment records also lacked documented consent and ethical approval or waiver. This is a limitation of the conduct of the study rather than a statistical issue. De-identification of the published presentation does not remove it. Prospective classroom investigations can integrate approved consent and data-use procedures into their design before research collection or secondary analysis begins.

6 Conclusion

The 26 students obtained higher recorded production and stress-perception scores after approximately eight weeks of explicit phonology instruction. Mean changes were 3.44 and 4.31 marks, respectively, and both paired comparisons were statistically significant; the production result held for each of the two raters separately. Production changes were more consistent, while the associations between outcomes remained uncertain. These results establish change in the recorded classroom assessments; they do not isolate a causal instructional effect or establish equivalent development across modalities. The study's contribution lies in documenting both productive and receptive performance within the same university course, with production scored independently by two raters, and in showing why their individual trajectories deserve separate attention. Interpretation remains bounded by test-form differences, rating limitations, and the disclosed conditions of research use.

Funding: Not stated in the submitted manuscript.

Conflicts of Interest: Not stated in the submitted manuscript.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

Abdulrahman, A. A., & Ramamoorthy, L. (2021). Erroneous word-stress patterns used by EFL Yemeni learners. *Indian Journal of Language and Linguistics, 2*(4), 39–48. https://doi.org/10.54392/ijll2145

Alzi'abi, S. E. (2025). Arab EFL learners' stress performance on suffixed words: Enhancing their lexical stress proficiency. *Language Testing in Asia, 15*, Article 76. https://doi.org/10.1186/s40468-025-00413-5

Alzi'abi, S. E. (2026). Phonological transfer and targeted instruction in Arab EFL learners' lexical stress acquisition. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688251396270

Baagbah, S. Y., & Ganapathy, M. (2025). Examining the use of PRAAT to improve English lexical stress production: Insights from self-directed learning within the framework of self-determination theory. *Language Teaching Research Quarterly, 49*, 170–193. https://doi.org/10.32038/ltrq.2025.49.09

Balla, A. A. S., & Elmahdi, O. E. H. (2024). Pronunciation challenges of Saudi female EFL learners: A multidimensional needs analysis. *British Journal of Applied Linguistics, 4(2)*, 15–30. https://doi.org/10.32996/bjal.2024.4.2.3

Chayluxsamee, Y., & Lynch, R. (2022). A comparative study of Grade 5 students' word stress competence under multisensory and listen and repeat teaching approaches in English as a foreign language class at a public school in Bangkok, Thailand. *Scholar: Human Sciences, 14*(1), 13–23.

https://assumptionjournal.au.edu/index.php/Scholar/article/view/4862

Chen, H. C., & Tian, J. X. (2024). The effects of metaphonological awareness protocol training on English word stress by Chinese EFL learners. *English Teaching & Learning, 48*, 697–724. https://doi.org/10.1007/s42321-023-00162-9

Chen, H. C., Han, Q., & Tian, J. X. (2025). The effects of rule-based and acoustic-perceptual instruction on Chinese ESL learners' English primary word stress placement accuracy. *International Journal of Applied Linguistics, 35*(3), 1149–1165. https://doi.org/10.1111/ijal.12687

Couper, G. (2022). A conceptual approach to teaching L2 pronunciation: Perception of word stress. *TESL-EJ, 26*(1). https://doi.org/10.55593/ej.26101a6

Deng, X., Wu, Z., & Chen, A. (2024). The effect of English songs on Chinese students' acquisition of word stress in a rural junior high school. *International Journal of Languages, Literature and Linguistics, 10*(2), 204–208.

Duris, M., Levis, J., Neiriz, R., & Silpachai, A. (2024). Methodological influences on word stress identification: Implications for research and teaching. In *Speech Prosody 2024* (pp. 732–736). https://doi.org/10.21437/SpeechProsody.2024-148

Elmahdi, O. E. H., & AbdAlgane, M. (2024). Teaching pronunciation to EFL learners at the tertiary level: Perspectives of professors at Saudi universities. *Journal of English Language Teaching and Applied Linguistics, 6(4)*, 86–95. https://doi.org/10.32996/jeltal.2024.6.4.9

https://doi.org/10.18178/IJLLL.2024.10.2.513

Helal, S. M. M., Mahdi, H. S., & Thabet, M. A. (2024). Captioning for clarity: Mastering stress with word-based vs. phonetic video transcriptions. *Forum for Linguistic Studies, 6*(4), 82–96. https://doi.org/10.30564/fls.v6i4.6779

Hirschi, K., & Kang, O. (2024). Data-driven learning for pronunciation: Perception and production of lexical stress and prominence in academic English. *TESOL Quarterly, 58*(3), 1081–1107. https://doi.org/10.1002/tesq.3302

Hosaka, I., & Saito, K. (2026). Dimension-specific roles of pitch and non-pitch processing in instructed L2 suprasegmental learning. *Bilingualism: Language and Cognition*. Advance online publication.

https://doi.org/10.1017/S1366728926101576

Ivanova, M. (2024). Training English word stress perception and production with technology. *The EUROCALL Review, 31*(2), 45–56. https://doi.org/10.4995/eurocall.2024.20400

Kanoksilapatham, B. (2025). "Teacher talk": An empowering resource for EMI teachers across disciplines. *Studies in English Language and Education, 12*(2), 613–631. https://doi.org/10.24815/siele.v12i2.41071

Kanoksilapatham, B., & Pariwatnanont, C. (2024). Unlocking potential: Integrating pronunciation into business instruction to inspire students and strengthen content teachers. *Studies in English Language and Education, 11*(2), 710–727. https://doi.org/10.24815/siele.v11i2.34767

Kanoksilapatham, B., & Takrudkaew, T. (2026). Enhancing TOEIC listening performance through explicit lexical stress instruction. *World Journal of English Language, 16*(6), 513. https://doi.org/10.5430/wjel.v16n6p513

Kim, K., & Archibald, J. (2024). The efficacy of lexical stress diacritics on the English comprehensibility and accentedness of Korean speakers. *Pedagogical Linguistics, 5*(1), 76–101. https://doi.org/10.1075/pl.22001.kim

Levis, J. M. (2018). Word stress and intelligibility. In *Intelligibility, oral communication, and the teaching of pronunciation* (pp. 100–124). Cambridge University Press. https://doi.org/10.1017/9781108241564.008

Maspufah. (2022). Utilizing YouTube video as media in teaching students' word stress pronunciation. *Pendekar: Jurnal Pendidikan Berkarakter, 5*(4), 23–28. https://journal.ummat.ac.id/index.php/pendekar/article/view/12287

Nakplad, R., Chantarangkul, V., Sermsook, K., & Eiadkaew, P. (2023). Using an animation movie to develop ability of stress in English of primary school students. *Education Quarterly Reviews, 6*(2), 118–123.

https://doi.org/10.31014/aior.1993.06.02.743

Roach, P. (2009). *English phonetics and phonology: A practical course* (4th ed.). Cambridge University Press.

https://www.cambridge.org/elt/peterroach/default.asp

Sain, Y., & Cobar, J. (2023). Improving students' pronunciation in word stress through TTS (Text to Speech) feature: On a Google Translate application. *IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, 11*(1), 837–844. https://doi.org/10.24256/ideas.v11i1.3321

Sardegna, V. G., & Jarosz, A. (2023). Learning English word stress with technology. In R. I. Thomson, T. M. Derwing, J. M. Levis, & K. Hiebert (Eds.), *Proceedings of the 13th Pronunciation in Second Language Learning and Teaching Conference*. https://doi.org/10.31274/psllt.16142

Sippel, L., & Martin, I. A. (2023). Immediate and long-term improvement in lexical stress perception: The role of teacher and peer feedback. *International Review of Applied Linguistics in Language Teaching, 61*(3), 1173–1195. https://doi.org/10.1515/iral-2021-0175

Thongsongsee, J., & Watanapokakul, S. (2023). An analysis of medical students' performance on the word stress patterns in English polysyllabic medical terms. *rEFLections, 30*(1), 12–37. https://doi.org/10.61508/refl.v30i1.263555

Yahy, R. A. (2022). The effect of explicit teaching of English stress rules on EFL students' performance at university level. *Adab Al-Rafidayn, 52*(89), 103–120. https://doi.org/10.33899/radab.2022.174387

Yang, Y., Zhang, Y., Zhou, Q., & Porter, D. A. (2026). Training English lexical stress with cue-signaling gestures for Chinese adolescents across different stress and intonation patterns. *Journal of Speech, Language, and Hearing Research, 69*(5), 1977–1992. https://doi.org/10.1044/2026_JSLHR-25-00527

Zhou, Y., Tang, Y., Zhu, J., & Chen, X. (2025). Sequence makes a difference: Effects of differential blended instructions on English word stress identification by Chinese teenager learners. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688251378445

Appendix A. Agreement between raters and sensitivity analysis

Table A1 Agreement between the two raters' production scores (N = 26)

Statistic	Pretest	Posttest	Change
Pearson r [95% CI]	.787 [.574, .900]	.862 [.712, .937]	.527 [.176, .760]
ICC(2,1) [95% CI]	.781 [.573, .895]	.865 [.722, .937]	.520 [.184, .750]
Quadratic-weighted kappa	.774	.860	.511
Mean difference (R1 − R2)	−0.31	0.12	0.42
SD of differences	1.16	1.03	1.65
95% limits of agreement	[−2.58, 1.96]	[−1.91, 2.14]	[−2.82, 3.66]
Mean absolute difference	1.08	0.96	1.42
Exact agreement (students)	1 of 26	2 of 26	6 of 26
Within ±1 mark (students)	24 of 26	25 of 26	10 of 26
Largest absolute difference (marks)	3	2	3
Bias test (one-sample t), p	.188	.574	.204

Note. R1 = first rater (course instructor); R2 = second rater. Scores are totals out of 20 and change = posttest − pretest. ICC(2,1) = two-way random-effects, absolute-agreement, single-rater intraclass correlation coefficient (F-based 95% CI). Quadratic-weighted kappa treats the integer totals as ordered categories. Limits of agreement = mean difference ± 1.96 SD of differences. Bias test = one-sample t-test of the differences against zero. Unadjusted descriptive statistics.

Table A2 Paired comparisons of production scores by rater

Scoring	Mean gain	SD gain	95% CI	t(25)	p	d_z
First rater (R1)	3.65	1.67	[2.98, 4.33]	11.14	< .001	2.19
Second rater (R2)	3.23	1.73	[2.53, 3.93]	9.54	< .001	1.87
Mean of raters (analysed)	3.44	1.49	[2.84, 4.04]	11.82	< .001	2.32

Note. N = 26. Change = posttest minus pretest, in marks out of 20. Pretest and posttest means (SD): R1 10.85 (1.80) and 14.50 (1.98); R2 11.15 (1.74) and 14.38 (1.94); mean of raters 11.00 (1.67) and 14.44 (1.89). Students with higher, unchanged, and lower scores: R1 25, 0, 1; R2 24, 2, 0; mean of raters 25, 0, 1. All two-sided p values are below .001; Wilcoxon signed-rank tests were also significant for each scoring (p < .001).