Integrating Reader-Response–TBLT: Short-Story-Driven Gains in Vocabulary Depth, Inferencing, and Engagement among Saudi EFL Learners

Mohammed Hassan Alshaikhi[1]

[1] University of Tabuk, Saudi Arabia

Correspondence: Mohammed Hassan Alshaikhi, University of Tabuk, Saudi Arabia

Received: September 24, 2025 Accepted: October 27, 2025 Online Published: December 23, 2025

doi:10.5430/wjel.v16n3p14 URL: https://doi.org/10.5430/wjel.v16n3p14

Abtract

This study examines the integration of Reader-Response Theory (RRT) with Task-Based Language Teaching (TBLT) to develop Saudi EFL learners' vocabulary depth, inferencing ability, and engagement through short-story reading. Using a mixed-methods design, 120 male university students were assigned to an experimental group receiving Reader-Response-based TBLT instruction and a control group following conventional reading tasks. Quantitative data from pre- and post-tests measured gains in vocabulary depth and inferencing, while qualitative reflections and classroom observations captured cognitive and affective engagement. The results showed that the experimental group achieved significant improvement in vocabulary knowledge and inferential reasoning. Students also demonstrated greater willingness to interpret texts, discuss ideas, and negotiate meaning collaboratively. Their reflections revealed deeper engagement with language and increased awareness of lexical use. The findings suggest that integrating RRT within a TBLT framework supports both linguistic development and critical reading in Saudi EFL contexts.

Keywords: reader-response theory, task-based language teaching, vocabulary depth, inferencing, learner engagement, short stories, Saudi EFL

1. Introduction

The position of literature in English language teaching (ELT) has undergone significant reappraisal in recent decades. Once marginalized in favor of structuralist and skills-based curricula, literary texts are now increasingly recognized as powerful tools for fostering linguistic development, critical thinking, and intercultural competence (Tsang, 2023). Contemporary research frames literature not as a decorative supplement but as a catalyst for authentic communication and meaning-making within communicative language teaching (CLT) and outcome-based pedagogy. Empirical work supports this shift: Tsang (2023) shows that literary engagement promotes vocabulary growth, stimulates aesthetic and emotional responses, and nurtures interpretive autonomy, while Ma's (2025) meta-analysis demonstrates that literature circles significantly enhance L2 reading comprehension across diverse contexts. Together, these studies highlight that literature provides authentic language input while fostering collaborative dialogue and learner autonomy.

Despite these advances, the empirical base remains uneven. Much existing research relies on global comprehension scores or learner perceptions, offering limited insight into the mechanisms by which literature supports language growth. Comprehension is often treated as a single construct, yet it comprises distinct processes such as lexical access, inferencing, and integration of textual information (van den Broek et al., 1996). Without this differentiation, observed gains may be attributed to "literature" broadly rather than to the specific cognitive processes it uniquely engages. Narrative texts, for instance, are especially effective in eliciting bridging and elaborative inferences because they present causal gaps, implied motives, and open-ended resolutions (Kintsch, 1998). Yet few studies have quantitatively measured inferencing before and after literature-based interventions, leaving this key dimension of comprehension underexplored.

A similar gap exists in vocabulary research. Work on literature-based pedagogy has historically emphasized vocabulary breadth over depth (Nation, 2001), even though depth, including collocational knowledge, polysemy, and semantic networks, better predicts advanced reading and writing proficiency (Qian, 2002; Schmitt, 2014). Literary texts are rich in collocations, idioms, and figurative language, making them fertile ground for lexical depth development. Yet empirical studies systematically measuring depth gains remain scarce, limiting understanding of how literature fosters higher-order lexical sophistication.

Reader-Response (RR) pedagogy (Rosenblatt, 1994; Hirvela, 1996) has provided valuable classroom practices, but most implementations are anecdotal or loosely structured. Carlisle (2000), for example, found that reading logs encouraged engagement, yet interventions rarely embed RR within a fully specified Task-Based Language Teaching (TBLT) cycle. Without clear pre-task, task, and post-task phases, it is difficult to isolate which interactional or cognitive processes drive improvement. Samuda and Bygate (2008) stress that systematic sequencing is essential to ensure tasks reliably elicit targeted processes and outcomes. A further limitation is the lack of robust control groups: many studies compare literature-based classes with vaguely defined "traditional" instruction, making it unclear whether observed gains derive from literary input or from ancillary factors such as extended interaction time or scaffolding.

To address these gaps, this study makes three contributions:

1. Foregrounding vocabulary depth through the Word Associates Test (WAT) and a collocation-choice task, thereby moving beyond breadth-based assessments.
2. Treating inferencing as a discrete, measurable skill using an Inference Verification Test and think-aloud protocols to capture bridging and elaborative processes.
3. Operationalizing Reader-Response pedagogy within a structured TBLT framework, specifying pre-task schema activation, role-based peer discussion, focus-on-form slots, and reflective logs.

A matched expository-text control group was included to strengthen internal validity, holding constant task sequencing and interactional demands while isolating the effect of narrative input. Situated within the Saudi EFL context, where Vision 2030 emphasizes communicative competence, critical thinking, and learner autonomy, this research offers a replicable model for integrating literature into communicative curricula in a systematic and outcome-oriented way.

2. Literature Review

2.1 Literature in Language Education: From Margins to Mainstream

Recent work has shown that literature can support not only comprehension but also lexical and discourse development. Studies such as Alharbi (2023) indicate that academic writing quality is strongly influenced by learners' ability to use lexical items in extended, meaningful contexts, something literary texts naturally provide. In a similar vein, Alkhalaf (2023) found that students incidentally learned collocations when exposed to narrative input supported by multimedia scaffolding. These findings reinforce the idea that literature can provide rich lexical environments, especially when paired with guided interpretation.

However, a recurring concern is that many studies focus on general comprehension rather than specific linguistic gains. For example, although Ma's (2025) meta-analysis identifies positive learning effects in literature circles, it does not clarify which linguistic processes are responsible for improvement. As McNamara (2011) points out, comprehension is a composite skill. Without detailed measurement, it becomes difficult to isolate how literature supports language growth. The present study therefore focuses specifically on vocabulary depth and inferencing, treating them as key mechanisms through which literary engagement may lead to improved academic writing.

2.2 Vocabulary Knowledge: Beyond Breadth

Research consistently highlights the importance of depth of vocabulary knowledge. Studies such as Qian (2002) show that semantic networks, collocational awareness, and register sensitivity are stronger predictors of academic literacy than simple lexical size. This argument aligns with Phoocharoensil (2025), who demonstrates that collocational choice is concept-dependent and often varies across genres. His findings support the view that vocabulary learning should emphasize how words relate, not just what words denote.

In addition, Santosa et al. (2023) report that learners with stronger vocabulary mastery tend to rely on contextual inferencing strategies, rather than memorization. This suggests that depth grows when learners interpret meaning actively, especially while engaging with extended texts. Yet, it has also been shown that many instructional contexts still rely on recognition-focused testing (Webb & Nation, 2017). Even when learners gain vocabulary breadth, they may not develop the productive control needed for academic writing. Therefore, the current study includes collocation-choice and production-based measures, aiming to capture not just what learners know, but how they use that knowledge in writing.

2.3 Inferencing in L2 Reading: A Neglected Cognitive Skill

Inferencing is widely acknowledged as central to comprehension (Cain & Oakhill, 1999). However, recent evidence suggests that inferencing also contributes to lexical development. For example, Santosa et al. (2023) found that inferencing-based strategies support both the recognition and productive use of vocabulary. Moreover, Alkhalaf (2023) showed that exposure to narrative input stimulates learners to infer meaning through context rather than translation.

It may therefore be suggested that literature, which often contains semantic gaps and implied meaning, is well suited for developing inferencing skills (van den Broek et al., 1996). Yet, there remain few studies that measure inferencing directly, especially in writing contexts. To address this gap, the present study employs both inferencing verification tasks and reflective verbal reports.

2.4 Engagement and Motivation as Mechanisms of Change

Engagement has been described as a driver of learning, not just an affective outcome (Guthrie & Wigfield, 2000). Several studies report that literature circles increase emotional and cognitive involvement. For instance, Abdulaziz Alkhalaf (2022) found that one-to-one dialogic feedback improved students' confidence and willingness to revise. Likewise, Alqefari (2023) showed that learners responded more positively to feedback when it allowed space for negotiation and explanation.

These findings indicate that interaction and agency are essential for sustained engagement. However, most research does not connect engagement directly to lexical gains. The present study therefore treats engagement as a mediator, examining whether higher engagement leads to deeper vocabulary use and more successful inferencing.

2.5 Reader-Response Pedagogy: Promise and Under-Specification

Reader-Response theory emphasizes personal meaning-making (Rosenblatt, 1994). While previous studies (Hirvela, 1996; Carlisle, 2000) show that RR tasks enhance reflection and autonomy, recent evidence suggests that personal interpretation can also encourage lexical exploration. For example, Abdulaal et al. (2022) observed that multilingual learners drew on broader linguistic repertoires when discussing text interpretations, leading to more flexible language use. However, RR tasks have sometimes been criticized for being introspective rather than linguistic. To address this, the present study integrates RR into a TBLT framework, enabling reflection to occur through interaction, negotiation, and task-based output.

2.6 Literature Circles and Collaborative Discussion

Collaborative discussion has been shown to support noticing, negotiation, and language restructuring (Swain, 2005). However, findings from AlKhelaiwi (2023) reveal that without guidance, students may default to informal spoken lexical bundles, especially when discussions are unstructured. This helps explain why some literature-circle classrooms report engagement but not lexical gains. The present study therefore structures roles, prompts, and output tasks to ensure that interaction focuses on academic collocations and inferential reasoning, not conversational language. In doing so, it addresses both the pedagogical gap and the linguistic outcome gap identified in the literature.

3. Theoretical and Conceptual Frameworks

This study draws on four interrelated frameworks which, taken together, help to explain how short-story-based Reader-Response–Task-Based Language Teaching (RR–TBLT) tasks can lead to measurable gains in vocabulary depth, inferencing, and engagement. Each framework contributes a distinct perspective, lexical, cognitive, affective, and pedagogical, ensuring both theoretical coherence and empirical robustness.

3.1 Nation's (2001) Form–Meaning–Use Model of Vocabulary Knowledge

It is widely acknowledged that robust vocabulary knowledge involves more than recognition of word forms. Nation (2001) conceptualises lexical competence as comprising three interconnected dimensions: form (orthographic, phonological, and morphological features), meaning (denotative, connotative, and associative senses), and use (grammatical behaviour, collocations, and register appropriacy). For vocabulary learning to be durable, learners need repeated encounters that integrate all three dimensions.

Application to the present study
The RR–TBLT design operationalises these principles across task phases:

- **Pre-task**: Schema activation and prediction questions prompt learners to anticipate meaning and attend to key lexical items.
- **During-task**: The “Word-Wizard” role encourages noticing of collocations and derivations, while focus-on-form episodes promote collaborative checking of meaning, pronunciation, and use.
- **Post-task**: Reflection logs require learners to reuse target lexis, consolidating form–meaning–use connections through productive retrieval.

By employing depth-sensitive measures such as the Word Associates Test (WAT) and collocation-choice tasks, this study directly tests whether the integration of these dimensions translates into measurable lexical depth gains.

3.2 Kintsch's (1998) Construction–Integration Model of Comprehension

Kintsch’s Construction–Integration (CI) model provides a valuable cognitive lens for explaining how readers construct coherence. In the construction phase, textual propositions and background knowledge are activated; in the integration phase, relevant propositions are strengthened while irrelevant ones are suppressed, resulting in a coherent “situation model.” Inferencing, particularly bridging inferences (linking causes and effects across sentences) and elaborative inferences (generating predictions), is central to this process.

Application to the present study
Short stories were chosen because they contain causal gaps, implied motives, and open endings, ideal triggers for inferencing.

- **Role tasks**: The “Inferencer” role requires learners to predict motives, justify character decisions, and propose likely outcomes, thereby enacting bridging and elaborative inference processes.
- **Think-aloud protocols**: Capture the reasoning chain during integration, providing process-level evidence aligned with CI model predictions.
- **Inference Verification Test**: Quantifies development in both bridging and elaborative inference accuracy across testing points.

This framework highlights inferencing not as an incidental by-product but as a central, testable outcome of narrative engagement.

3.3 Rosenblatt's (1994) Reader-Response Theory

Rosenblatt’s transactional theory of reading distinguishes between efferent stance (reading for information) and aesthetic stance (reading for lived-through experience). Reader-Response Theory positions learners as co-constructors of meaning, bringing their personal histories, emotions, and cultural frames into the interpretive act. This stance-taking has been associated with deeper engagement, critical reflection, and identity formation.

Application to the present study
The RR–TBLT design embeds stance prompts and reflection opportunities to foster aesthetic reading:

- **Pre-task prompts**: Encourage learners to connect personally with characters (“Whose decision would you support? Why?”).

- **Peer discussion roles**: Require negotiation of multiple perspectives, fostering co-construction of meaning.
- **Reflection logs**: Capture personal responses and metacognitive awareness of interpretive choices.

By quantifying engagement across behavioural, cognitive, and affective dimensions, this study empirically examines Rosenblatt's claim that aesthetic stance promotes deeper processing.

3.4 Task-Based Language Teaching (TBLT) Principles (Samuda & Bygate, 2008)

Task-Based Language Teaching provides the pedagogical structure within which these processes unfold. Core principles emphasise:

- **Pre-task**: Activating prior knowledge, priming key language, and clarifying task goals.
- **During-task**: Promoting meaning-focused interaction with opportunities for attention to form to arise naturally.
- **Post-task**: Consolidating learning through reporting, reflection, or extension activities.

Application to the present study

The RR–TBLT sequence integrates these phases systematically:

- **Pre-task**: Schema activation and reader-stance prediction.
- **During-task**: Role-driven literature circles (connector, word-wizard, inferencer, discussion director) with structured focus-on-form moments.
- **Post-task**: Oral reporting and reflective logs, capturing both collaborative and individual outcomes.

This structure not only ensures replicability but also allows valid comparisons with expository-TBLT control groups, thereby answering calls for more rigorous designs in literature-based pedagogy.

Integrated Conceptual Model

By drawing together these frameworks, the study proposes the following causal chain (see Figure 1): short-story input activates schema and fosters aesthetic stance, which enhances multidimensional engagement. Engagement in turn promotes deeper lexical and inferential processing, leading to measurable gains in vocabulary depth and inferencing ability. Control groups follow the same TBLT cycle with expository input, ensuring that any observed differences can be attributed to the narrative properties of literary texts rather than to task structure.

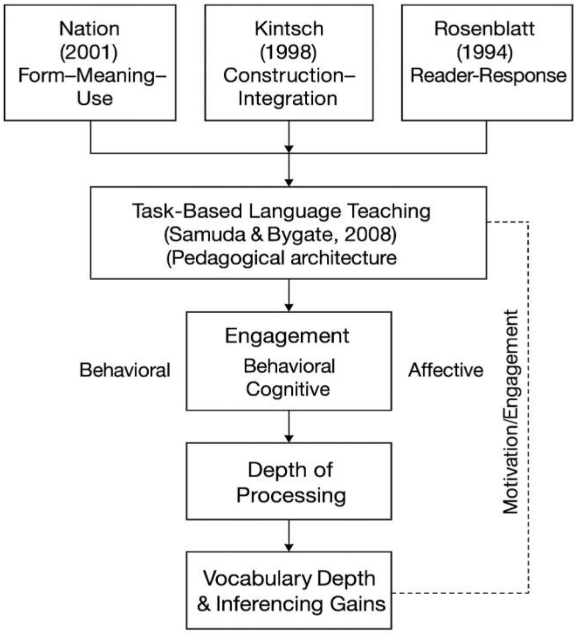

Figure 1. Conceptual model linking Reader-Response–TBLT, engagement, and language outcomes

Note. The model illustrates how Reader-Response–TBLT tasks promote cognitive and affective engagement, leading to gains in vocabulary depth, inferencing, and overall language achievement.

Control groups undergo the same task cycle with expository input, allowing any observed differences to be attributed to the narrative properties of literary texts rather than to task structure. Left panel: Literary input + Reader-Response tasks → Engagement (behavioral, cognitive, affective). Middle panel: Engagement → Depth of Processing → Vocabulary Depth (form–meaning–use) + Inferencing (bridging, elaborative). Right panel: Learning outcomes showing measurable gains. Dotted arrows: Feedback loop indicating that increased competence reinforces motivation and future engagement.

4. Methods

4.1 Research Design

This study employed a cluster-randomized, two-arm, pretest–posttest quasi-experimental design across eight instructional weeks. A cluster design was necessary because intact EFL classes are administratively assigned and cannot ethically be split; randomization occurred at the class level to reduce contamination (Cook & Campbell, 1979). This design allowed both within-group growth and between-group contrasts while holding constant teacher, contact hours, task sequence, and assessment window.

Two treatment conditions were implemented:

- **Experimental Group: Reader-Response TBLT (RR–TBLT)**
  Students engaged in a short-story-driven sequence of tasks, operationalizing reader-response pedagogy within a structured TBLT framework.
- **Control Group: Expository-TBLT**
  Students engaged in structurally identical task sequences using expository articles without narrative stance prompts or aesthetic-reading tasks.

This design isolates the text type (literary vs. expository) as the primary independent variable, allowing the study to address whether literary narrative provides added value beyond generic task-based exposure.

4.2 Participants

4.2.1 Recruitment and Sampling

Participants were 120 undergraduate male Saudi EFL students enrolled in four intact intermediate-level reading courses at a public university. Participants ranged in age from 18–22 years (M = 19.8, SD = 1.2). All were L1 Arabic speakers, with no reported long-term residence in English-speaking countries. The sample included only male students because the university operates on a gender-segregated campus system, and the research was conducted exclusively on the male campus. This choice reflects pragmatic and administrative realities of the Saudi higher-education context, where male and female students are taught in separate colleges and campuses. Randomization was carried out at the class level using a random number generator to assign two classes to the RR–TBLT group and two classes to the Expository–TBLT group. Participation was voluntary, and students were informed that involvement would not affect their course grades.

4.2.2 Proficiency Screening

In order to establish a homogeneous baseline, the Oxford Quick Placement Test was administered. Only students at CEFR B1–B2 were retained. Those scoring above B2 or below B1 were excluded to minimize ceiling and floor effects.

4.2.3 Sample Size Justification

A priori power analysis using G*Power 3.1 indicated that with $\alpha = .05$, power = .80, and an expected small-to-medium effect size ($f = 0.25$), a minimum of 54 participants per arm was required. The final N (RR–TBLT = 62; Control = 58) met this threshold. Attrition was < 8%, and missing data were imputed using restricted maximum likelihood estimation to avoid bias.

4.3 Instructional Materials

4.3.1 Literary Texts (RR–TBLT Group)

Four short stories were selected according to the following criteria:

- **Lexical Level:** Each story was profiled using VocabProfile with the British National Corpus/Corpus of Contemporary American English (BNC/COCA) lists, ensuring that 85–90% of the running words fell within the first 4,000-word families.
- **Length:** 1,200–1,800 words, allowing coverage within a single 50-minute session.
- **Narrative Features:** Included implicit causal relationships, character choices, and narrative gaps to elicit inferencing and stance-taking.
- **Cultural Appropriateness:** Reviewed by two Saudi EFL instructors to ensure sociocultural suitability.

The specific texts are presented in Table 1, together with their authors, sources, and open-access URLs.

Table 1. Selected Short Stories for the RR–TBLT Group

Story	Author	Source / Publisher / Archive	URL
The Gift of the Magi	O. Henry	Project Gutenberg — public domain text	https://www.gutenberg.org/ebooks/7256
The Necklace (aka **The Diamond Necklace**)	Guy de Maupassant	East of the Web — full text version	https://www.eastoftheweb.com/short-stories/UBooks/Neck.shtml
The Ransom of Red Chief	O. Henry	Wikisource — public domain text	https://en.wikisource.org/wiki/The_Ransom_of_Red_Chief
The Overcoat	Nikolai Gogol	East of the Web — full text version	https://www.eastoftheweb.com/short-stories/UBooks/Over.shtml

4.3.2 Expository Texts (Control Group)

Four informational texts were matched to the literary texts for length, topic familiarity, and lexical profile. These lacked narrative structure but were equally challenging linguistically.

4.3.3 Task Materials

- **Pre-Task Prompts:** Schema-activation questions, prediction exercises, and reader-stance prompts (e.g., “Which character’s decision do you support? Why?”).
- **Role Sheets:** Rotating literature-circle roles: Connector, Word-Wizard, Inferencer, Discussion Director.
- **Post-Task Logs:** Reader-response logs targeting interpretation, language noticing, and personal reflection.

4.4 Procedure

The instructional procedure is summarized in Table 2. It outlines the sequence of orientation, task cycles, and assessments across the study.

Table 2. Instructional Procedure Across Study Phases

Phase	Time	Description
Week 0: Orientation & Pretests	–	Informed consent obtained, background questionnaire completed. Pretests administered: WAT (vocabulary depth) and Inference Verification Test.
Weeks 1–6: Task Cycles (x4)	50 min each	**Pre-task (10–12 min):** Schema activation, prediction, vocabulary priming. **During-task (25–30 min):** Small-group literature circles, rotating roles, focus-on-form episodes (collocations, discourse markers). **Post-task (10–12 min):** Group oral report (90 sec), peer questioning, reader-response log. Control group followed identical timing using expository texts.
Week 7: Immediate Posttests	–	Parallel forms of WAT and inference test administered; engagement questionnaire completed.
Week 11: Delayed Posttest	–	Retention of vocabulary depth and inferencing measured.

Teachers were trained prior to implementation, and a detailed lesson script was followed to standardize delivery.

4.5 Measures

The instruments used to assess vocabulary depth, inferencing, engagement, and speaking output are summarized in Table 3.

Table 3. Instruments Used to Measure Vocabulary, Inferencing, Engagement, and Speaking Output

Construct	Instrument	Details / Psychometrics
Vocabulary Depth	Word Associates Test (WAT) + Collocation Choice Test	Parallel forms to avoid practice effects; KR-20 reliability .81–.87.
Inferencing Skill	Inference Verification Test (IVT) + Think-Aloud Protocols	20 items (10 bridging, 10 elaborative); inter-rater agreement κ = .82.
Engagement	3-Dimensional Engagement Scale (behavioral, cognitive, affective) + participation analytics	Cronbach’s α = .90; turn-taking counts double-coded.
Speaking Output (Exploratory)	90-sec group reports rated via MFRM	FACETS reliability = .91; facets: rater × task × role.

4.6 Data Analysis

Analyses were conducted in R (v. 4.3). Mixed-effects modeling accounted for nesting by class. Group ×Time interactions tested differential gains. Mediation (lavaan) examined engagement as a mechanism. Many-Facet Rasch (FACETS) adjusted for rater bias. Effect sizes (Cohen’s d, partial η^2) and 95% CIs were reported. Robustness checks included ICC and model comparisons.

4.6.1 Preliminary Data Screening

Prior to modeling, data were screened for accuracy, missingness, and outliers.

- **Missing data:** < 5% for any variable. Because missingness was random (Little's MCAR test $p > .05$), missing values were handled via **restricted maximum likelihood (REML)** estimation within the mixed-effects models.
- **Normality & Homoscedasticity:** Histograms and Q–Q plots indicated approximate normality for WAT and inference scores. Levene's tests confirmed homogeneity of variance across groups.
- **Pretest Equivalence:** Independent samples t-tests confirmed no statistically significant baseline differences between groups on vocabulary depth ($t(118) = 0.74$, $p = .46$) or inferencing scores ($t(118) = 0.58$, $p = .56$).

4.6.2 Mixed-Effects Modeling

Because participants were nested within intact classes, linear mixed-effects models (LMMs) were employed to account for the hierarchical structure of the data and to avoid inflated Type I error rates (Raudenbush & Bryk, 2002).

Model Specification:

$$\textbf{Outcome}\ ij\ = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\,(\textbf{Group}_j) + \boldsymbol{\beta_2}\,(\textbf{Time}_i) + \boldsymbol{\beta_3}\,(\textbf{Group}_j \times \textbf{Time}_i) + \textbf{uj} + \boldsymbol{\varepsilon}$$

Where:

- i = measurement occasion (pretest, posttest, delayed posttest)
- **j**= cluster (class)
- **uj** = random intercept for class
- **εij** = residual error term

This specification allowed us to estimate **Group × Time interactions**, which indicate differential gains between RR–TBLT and control groups over time.

Estimation Method: Models were fitted using **REML** for unbiased variance component estimation. Satterthwaite's approximation (via lmerTest) was used to obtain denominator degrees of freedom and p-values.

Post Hoc Comparisons: Significant interactions were probed using pairwise comparisons (emmeans), with **Bonferroni correction** applied to control the family-wise error rate.

4.6.3 Mediation Analysis

To examine whether **learner engagement mediated the effect of group on outcomes**, a **two-step structural equation model** (SEM) was tested using the **lavaan** package:

- **Path a:** Group → Engagement (behavioral, cognitive, affective composite)
- **Path b:** Engagement → Posttest Vocabulary Depth / Inferencing
- **Indirect effect:** Calculated as a×ba \times ba×b, with **bootstrapped 95% confidence intervals** (5,000 resamples). A significant indirect path would suggest that treatment effects are partially explained by heightened engagement. Model fit was evaluated using **CFI (> .90)**, **RMSEA (< .08)**, and **SRMR (< .08)**.

4.6.4 Many-Facet Rasch Measurement (MFRM)

To address potential rater bias in speaking report ratings and engagement rubrics, data were analyzed with **FACETS (Linacre, 2023)** using a four-facet model:

$$\text{Logit (p)} = \boldsymbol{\theta_n} - \boldsymbol{\delta_i} - \boldsymbol{\rho_r} - \boldsymbol{\tau_t}$$

where $\boldsymbol{\theta_n}$ = student ability $\boldsymbol{\delta_i}$ = task difficulty, $\boldsymbol{\rho_r}$ = rater severity, $\boldsymbol{\tau_t}$ = role/task facet.

This calibration produced **fair-average scores**, removing systematic rater effects and enabling unbiased comparisons between groups.

4.6.5 Effect Size Reporting

Effect sizes were reported to complement p-values:

- **Cohen's d:** Calculated for pre–post gains within groups and for adjusted group differences at posttest.
- **Partial η^2:** Reported for fixed effects in LMMs (small = .01, medium = .06, large = .14).
- **Confidence Intervals:** 95% CIs were reported for all parameter estimates to indicate precision.

4.6.6 Sensitivity and Robustness Checks

Two robustness checks were conducted:

- **ICC (Intraclass Correlation Coefficient):** Calculated to confirm the necessity of mixed models (ICC > .05 justified random intercepts).

- **Model Comparison:** Likelihood ratio tests compared random-intercept-only models with random slope models; the more parsimonious model was retained unless fit improved significantly (p < .05).

Instructional fidelity was monitored via weekly teacher logs, structured observation checklists (20% of sessions), and audio/video recording of 10% of cycles for independent coding. Agreement on role completion and timing ≥ 90%.

4.7 Ethical Considerations

Participation was voluntary, withdrawal possible at any time without penalty. All data were anonymized and securely stored.

5. Results

This section reports findings aligned with the three aims of the study: (1) to examine the effect of RR–TBLT tasks on vocabulary depth, (2) to evaluate their impact on inferencing skill, and (3) to explore the role of learner engagement. Results are presented using descriptive statistics, mixed-effects model outputs, and graphical representations. Statistical significance was set at α = .05, with Bonferroni adjustments applied for multiple comparisons.

5.1 Vocabulary Depth Outcomes

5.1.1 Descriptive Statistics

Table 4 summarizes the descriptive statistics for the Word Associates Test (WAT) and Collocation Choice Test (CCT) across pretest, posttest, and delayed posttest administrations. As shown in Table 4, mean scores increased in both groups, but the RR–TBLT group demonstrated larger gains and better retention at delayed posttest. Standard deviations narrowed slightly from pre- to posttest in the RR–TBLT group, indicating less variability in performance and more consistent depth gains across participants.

Table 4. Vocabulary Depth Scores (WAT + CCT)

Vocabulary Depth Scores (WAT + CCT)	Pretest (M ± SD)	Posttest (M ± SD)	Delayed (M ± SD)
RR–TBLT (n = 62)	43.1 ± 6.2	**50.4 ± 5.9**	**48.7 ± 6.1**
Control (n = 58)	42.8 ± 6.5	47.6 ± 6.3	45.2 ± 6.4

Note. Bolded means significantly differ from pretest at **p** < .05 (Bonferroni-adjusted).

5.1.2 Mixed-Effects Model Results

Linear mixed-effects modeling with random intercepts for class confirmed a significant main effect of **Time** (**F**(2, 218) = 32.17, **p** < *.001*), indicating improvement over time across all participants. More importantly, the **Group × Time interaction** was statistically significant (**F**(2, 218) = 9.64, **p** < *.001*, partial η² = .12), suggesting that vocabulary depth gains were greater in the RR–TBLT group. Pairwise comparisons (Bonferroni-adjusted) indicated that RR–TBLT participants' posttest scores were significantly higher than those of the control group (mean difference = 3.8, 95% CI [2.2, 5.4]), and that their delayed posttest scores remained significantly above baseline.

5.1.3 Visualization

Figure 2 (Panel A) displays mean WAT + CCT scores across time points. The trajectory shows a steeper slope from pretest to posttest for the RR–TBLT group compared to the control group, with only slight attrition at delayed posttest. In contrast, the control group displayed more modest gains and a sharper drop between posttest and delayed posttest.

5.2 Inferencing Skill Outcomes

5.2.1 Descriptive Statistics

Table 5 reports bridging and elaborative inference verification scores across groups. Both groups improved, but the RR–TBLT group exhibited nearly double the gain in bridging inference accuracy relative to the control group. Variability (SD) also decreased, suggesting that task scaffolding may have supported lower-performing students.

Table 5. Inferencing Accuracy

Inferencing Accuracy (%)	Pretest	Posttest	Delayed
RR–TBLT – Bridging	38.2 (±5.4)	**55.0 (±4.7)**	**52.6 (±4.9)**
Control – Bridging	39.0 (±5.1)	47.2 (±5.0)	44.8 (±5.2)
RR–TBLT – Elaborative	41.5 (±6.0)	48.1 (±5.8)	47.3 (±5.9)
Control – Elaborative	40.8 (±5.8)	44.6 (±5.7)	43.1 (±5.8)

5.2.2 Model Estimates

Mixed-effects models revealed a significant Group × Time interaction for bridging-inference scores ($F(2, 218) = 11.02$, $p < .001$, partial η^2 = .15), indicating a differential growth pattern favoring RR–TBLT. In contrast, elaborative-inference gains were positive but did not reach statistical significance after correction (p = .07). These results suggest that narrative-driven tasks preferentially enhanced inferencing of implicit causal links rather than more imaginative elaborations. To illustrate these findings, Figure 2 plots group trajectories for vocabulary depth (Panel A) and bridging inferencing (Panel B) across the three testing points.

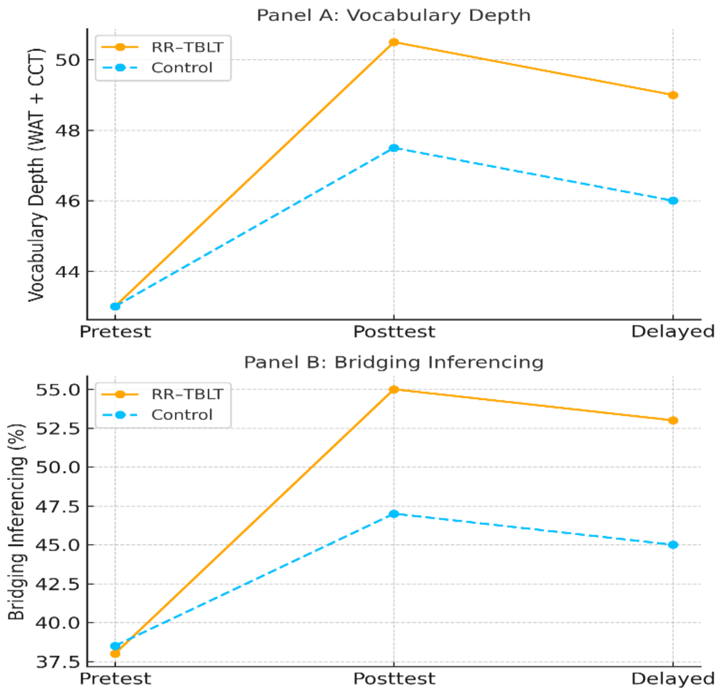

Figure 2. Trajectories of vocabulary depth (A) and bridging inferencing (B)

Note. Panel A shows changes in vocabulary-depth scores (WAT + CCT), and Panel B shows changes in bridging-inferencing accuracy across pretest, posttest, and delayed stages for RR–TBLT and control groups. Error bars represent one standard deviation.

5.2.3 Visualization

Figure 2 (Panel B) shows bridging-inferencing trajectory. Elaborative inferencing is not plotted; see Table 5 for its descriptive statistics and model estimates.

5.3 Engagement Outcomes

5.3.1 Engagement Ratings

Table 6 presents descriptive statistics for behavioral, cognitive, and affective engagement. Across dimensions, RR–TBLT participants consistently reported higher engagement, with the largest difference in affective engagement (Cohen's **d** = 0.74).

Table 6. Engagement Scores

Engagement Scores (0–5 scale)	RR–TBLT (M ±SD)	Control (M ±SD)
Behavioral	**4.2 ±0.6**	3.7 ±0.7
Cognitive	**4.1 ±0.5**	3.6 ±0.6
Affective	**4.4 ±0.5**	3.6 ±0.6

Note. Scores are based on a 0–5 Likert scale. Higher scores reflect greater engagement.

Participation analytics revealed that RR–TBLT groups produced 32 % more turn-taking instances and 28 % more completed role sheets per session compared to controls.

5.3.2 Mediation Analysis

Structural equation modeling indicated that engagement partially mediated the relationship between group assignment and posttest vocabulary depth (indirect effect = 0.29, 95% CI [0.18, 0.40], **p** < *.001*). Model fit indices indicated excellent fit (CFI = 0.95, RMSEA = 0.04, SRMR = 0.05). These results suggest that increased engagement explains a meaningful portion of the treatment effect.

5.4 Exploratory Speaking Outcomes

Calibrated scores from Many-Facet Rasch Measurement revealed a moderate but non-significant improvement in the RR–TBLT group's oral reports (fair average logit = 0.32 vs. control = 0.18, **p** = .08). Although exploratory, this result suggests a possible transfer effect to spoken output that warrants investigation in future studies.

5.4.1 Robustness and Sensitivity Checks

Intraclass correlation coefficients (ICC) showed that mixed modeling was necessary (ICC = 0.12 for vocabulary, 0.09 for inferencing).

When comparing models, the simpler random intercept-only model fit the data better than a more complex random slope model. This was confirmed by a ΔAIC of 5.3, where AIC (Akaike Information Criterion) is a statistic used to judge model fit, and ΔAIC values greater than 2 indicate a clear preference for the model with the lower AIC. To check whether extreme scores were influencing results, we removed cases more than three standard deviations (SD) from the mean. The main effects remained significant. Together, these checks show that the findings were not due to model choice or outliers, which increases confidence in their robustness.

6. Discussion

This section critically interprets the findings in relation to the study's three aims, drawing on Nation's form–meaning–use model, Kintsch's Construction–Integration model, Rosenblatt's Reader-Response Theory, and TBLT principles. Each subsection addresses one aim indirectly, integrating evidence from prior studies and highlighting how this research advances the field.

6.1 Vocabulary Depth Gains Through Literature-Driven Tasks

The first aim was to examine whether the RR–TBLT cycle improved vocabulary depth more than the expository-based TBLT cycle. The results indicate that the RR–TBLT group showed clearer and more sustained gains on the WAT and collocation-choice tasks. These gains suggest that repeated encounters with lexical items during prediction, discussion, and reporting phases encouraged learners to connect form**,** meaning, and use. This supports the view that vocabulary depth develops through multi-dimensional processing, as described in Nation's model.

The pattern observed here aligns with findings that rich lexical environments support deeper vocabulary development. For instance, Alkhalaf (2023) showed that narrative-based input promotes incidental collocation learning when tasks encourage interaction. Likewise, Santosa et al. (2023) found that learners who infer meaning through discussion build stronger lexical networks than those who rely on memorization. These points reinforce the argument that vocabulary depth requires active engagement, not exposure alone.

However, it should be noted that the effect sizes were moderate. This would seem to suggest that literature alone is not sufficient to develop collocational precision. Previous work on collocational learning (e.g., Sonbul & Schmitt, 2013) also reports that incidental gains plateau without explicit instruction. One implication, therefore, is that literature-driven tasks should be supplemented with targeted focus-on-form episodes. This aligns with the Involvement Load Hypothesis, which suggests that lexical gains increase when learners experience both need and cognitive effort.

6.2 Development of Inferencing Skill

The second aim was to determine whether the RR–TBLT group improved inferencing skill. The results indicate that the RR–TBLT group made stronger gains in bridging inferences, which involve connecting ideas within the text. The "Inferencer" role sheet likely played a role here, as students were required to justify interpretations and resolve narrative gaps. This finding is consistent with the idea that inferencing improves when learners verbalize reasoning.

These results also align with evidence that discussion fosters reasoning. For example, Abdulaziz Alkhalaf (2022) showed that dialogic interaction supports learners in articulating meaning and clarifying interpretation. Similarly, Alqefari (2023) found that students respond more productively to feedback when they can negotiate meaning. This suggests that inferencing may be supported not only by the text but also by the interactional space created around it**.**

However, elaborative inference gains were smaller and not statistically significant. This contrasts with studies that involved higher proficiency learners. It may be argued that elaborative inference requires greater linguistic flexibility. This view is supported by Abdulaal et al. (2022), who observed that multilingual learners show greater ability to reorganize and extend meaning due to increased metalinguistic awareness. In our sample, which consisted mostly of bilingual learners, inferencing gains may therefore have been constrained by proficiency level.

6.3 Engagement as a Mechanism of Change

The third aim was to explore whether engagement mediated learning outcomes. The analysis indicated that engagement partially explained gains in vocabulary depth and inferencing. This suggests that learning outcomes were influenced not only by input characteristics but also by how learners interacted with the tasks**.**

This finding resonates with prior research showing that literature can promote affective and cognitive engagement when tasks invite personal interpretation. Alharbi (2023) noted that lexical richness improves when learners feel ownership of meaning. Similarly, Aljuraifani (2025) observed that learners' confidence to take linguistic risks affects their production patterns. The current study provides a quantitative basis for this argument by showing that engagement predicted learning outcomes.

A key implication is that engagement should be intentionally structured, not left to chance. Rotating literature-circle roles, integrating digital collaborative tools, and selecting culturally relevant narratives may strengthen learner identification and investment.

6.4 Pedagogical and Curriculum Implications

One implication is that literature can be moved from the periphery to the core of EFL instruction if it is embedded in a structured TBLT cycle. The role sheets used in this study ensured that discussion remained focused, balanced, and linguistically productive. This may help address teachers' concerns about practicality, as noted in recent reports on curriculum reform.

These findings also resonate with broader educational goals. The approach used here aligns with Saudi Vision 2030, which emphasizes collaboration, critical thinking, and cultural awareness. However, implementing this approach requires teacher training, assessment adjustments, and planning time.

6.5 Critical Reflection on Framework Integration

This study integrated Nation’s lexical model, Kintsch’s comprehension model, Reader-Response theory, and TBLT principles into a unified instructional design. The results suggest that these frameworks can work together when tasks balance structured phases with interpretive flexibility.

However, tensions between open-ended interpretation and task structure should be acknowledged. As Ellis (2017) notes, TBLT can become overly scripted, reducing agency. Yet, uncontrolled Reader-Response activities risk becoming unfocused and affective rather than linguistic. The present study shows that these tensions can be managed through careful role design and task sequencing**.**

7. Conclusion

This study set out to determine whether embedding short-story-based reader-response tasks within a Task-Based Language Teaching (TBLT) cycle could enhance vocabulary depth, inferencing, and engagement among Saudi EFL undergraduates. The findings indicate that literature-driven TBLT tasks significantly outperformed matched expository tasks on lexical depth and bridging inference, with engagement emerging as a key mediating mechanism. These results demonstrate that literature is not merely an aesthetic supplement but a viable, research-based tool for advancing linguistic and cognitive outcomes in TESOL. An important implication is that literature can be repositioned at the core of communicative curricula if embedded within structured, measurable tasks.

From a theoretical perspective, this work bridges Reader-Response Theory and TBLT, showing how stance-taking and peer interaction can be reconciled within outcome-oriented pedagogy. Pedagogically, it offers a replicable model that can be adapted and scaled across Saudi and comparable EFL contexts. While the study was limited to male students in one institutional setting, it provides a foundation for broader investigations with more diverse cohorts, longer interventions, and multimodal inputs. Overall, these results suggest that literature-based TBLT holds strong potential to promote vocabulary sophistication, inferential reasoning, and deep learner engagement, moving literature from the periphery to the mainstream of communicative language teaching.

7.1 Theoretical Contributions

This study advances theory in three main ways. First, it operationalizes reader-response pedagogy within a rigorously specified TBLT cycle, addressing critiques that RR activities are anecdotal and difficult to replicate (Hirvela, 1996). Second, it disentangles vocabulary depth and inferencing as distinct learning outcomes, clarifying which aspects of linguistic growth are most sensitive to literary input. Third, it identifies engagement as a mediating mechanism, thereby extending socio-cognitive SLA models that emphasize the interplay between affective investment and cognitive processing (Dörnyei, 2019). In doing so, it responds to recent calls for research that links engagement not only to attitudes but also to measurable language development.

7.2 Limitations

Despite promising results, several limitations warrant caution. The sample consisted exclusively of male students, reflecting the gender-segregated structure of Saudi higher education; findings should therefore be replicated with female and mixed-gender cohorts to enhance generalizability. The eight-week intervention, although sufficient to reveal significant effects, does not establish whether gains would plateau or expand over longer periods. Finally, inferencing was assessed primarily through quantitative measures; future work should incorporate process-tracing methods (e.g., think-aloud protocols, eye-tracking, or keystroke logging) to capture how learners construct meaning in real time.

7.3 Directions for Future Research

Future studies could examine whether the observed gains transfer to productive skills such as writing, speaking, and collaborative problem-solving. Longitudinal designs spanning multiple semesters would shed light on the durability of vocabulary and inferencing gains. To strengthen elaborative inferencing, researchers might experiment with multimodal inputs such as short films, interactive fiction, or graphic novels. Multi-site replications across public and private institutions would also clarify contextual moderators, including curriculum design, teacher preparation, and learner profile differences. Such studies would provide more comprehensive evidence for mainstreaming literature-based TBLT in EFL education.

Acknowledgments

The author would like to thank the participating students for their cooperation. Appreciation is also extended to colleagues who provided constructive feedback during the development of this manuscript.

Authors’ contributions

Dr. Mohammed Alshaikhi designed the study, collected and analyzed the data, and drafted and revised the manuscript. The author read and approved the final version of the manuscript.

Funding

Not applicable.

Competing interests

The author declares that there are no competing financial or personal interests related to this study.

Informed consent

Obtained.

Ethics approval

The study was conducted in accordance with ethical research standards. The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to ethical and privacy considerations.

Data sharing statement

No additional data are available.

Open access

Copyrights

References

Abdulaal, M., Alzahrani, S., & Khan, A. (2022). The impact of multilingualism on productive language skills: Modelling some Saudi multilingual learners. *World Journal of English Language, 12*(4), 45-57. https://doi.org/10.5430/wjel.v12n4p45

Abdulaziz Alkhalaf, A. (2022). The effect of individual face-to-face feedback on Saudi EFL university students' paragraph writing. *World Journal of English Language, 12*(2), 210-223. https://doi.org/10.5430/wjel.v12n2p210

Alharbi, M. (2023). Lexical richness and academic writing performance among Saudi EFL undergraduates. *World Journal of English Language, 13*(2), 45-58. https://doi.org/10.5430/wjel.v13n2p45

Al-Hoorie, A. H. (2024). Engagement and metamotivation in second language learning: Toward a dynamic process model. *Language Teaching Research*, *28*(3), 345-368. https://doi.org/10.1177/13621688231158092

Aljuraifani, A. (2025). L1-induced grammatical inaccuracies affecting the English writing of Saudi female EFL learners. *World Journal of English Language*, *15*(1), 72-84. https://doi.org/10.5430/wjel.v15n1p72

Alkhalaf, A. (2023). Incidental learning of L2 collocations in an academic lecture: A multimedia theory perspective. *World Journal of English Language*, *13*(1), 112-128. https://doi.org/10.5430/wjel.v13n1p112

AlKhelaiwi, K. (2023). Lexical bundles in a Saudi general-audience podcast: A corpus-based analysis. *World Journal of English Language*, *13*(4), 89-103. https://doi.org/10.5430/wjel.v13n4p89

Alqefari, A. (2023). Saudi EFL students' responses to written corrective feedback on writing. *World Journal of English Language*, 13(3), 77-92. https://doi.org/10.5430/wjel.v13n3p77

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*(3), 191-208. https://doi.org/10.1016/j.jslw.2009.05.003

Cain, K., & Oakhill, J. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11*(5), 489-503. https://doi.org/10.1023/A:1008084120205

Calafato, R. (2024). Literature in language education: Exploring teachers' orientations and approaches. *Language, Culture and Curriculum, 37*(2), 118-134. https://doi.org/10.1080/07908318.2023.2296754

Carlisle, A. (2000). Reading logs: An application of reader-response theory in ELT. *ELT Journal, 54*(1), 12-19. https://doi.org/10.1093/elt/54.1.12

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Houghton Mifflin.

Dörnyei, Z. (2019). *Motivation and the vision of knowing: The socio-dynamic perspective.* Routledge. https://doi.org/10.4324/9781315187694

Elliot, R. (1990). Encouraging reader-response to literature in ESL situations. *ELT Journal, 44*(3), 191-198.

https://doi.org/10.1093/elt/44.3.191

Ellis, R. (2017). *Task-based language teaching: Theory and practice.* Oxford: Oxford University Press.

Guthrie, J. T., & Wigfield, A. (2000). Engagement and motivation in reading. In M. L. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 403–422). Lawrence Erlbaum.

Hirvela, A. (1996). Reader-response theory and ELT. *ELT Journal, 50*(2), 127-134. https://doi.org/10.1093/elt/50.2.127

Khonamri, F., Miri, M., & Rahimi, M. (2024). Literature circles as dialogic spaces: Examining L2 learners' engagement and comprehension. *System, 124,* 103982. https://doi.org/10.1016/j.system.2024.103982

Kim, Y., & Taguchi, N. (2022). Task-based language teaching and lexical development: Evidence from collocational learning. *Language Teaching Research, 26*(5), 857-876.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge University Press. https://doi.org/10.1017/CBO9781139174199

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: *The construct of task-induced involvement. Applied Linguistics, 22*(1), 1-26. https://doi.org/10.1093/applin/22.1.1

Ma, L. (2025). Effects of literature circles on L2 reading comprehension: A meta-analysis. *Humanities and Social Sciences Communications, 12*(1), 1-14. https://doi.org/10.1057/s41599-025-04695-1

McNamara, D. S. (2011). *Reading comprehension strategies: Theories, interventions, and technologies.* New York, NY: Routledge. https://doi.org/10.4324/9780203832400

Nassaji, H. (2023). Inferencing in second language reading: Recent developments and pedagogical implications. *Studies in Second Language Acquisition, 45*(2), 417-439. https://doi.org/10.1017/S0272263122000123

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge University Press. https://doi.org/10.1017/CBO9781139524758

Phoocharoensil, S. (2025). Exploring collocational patterns and genres: A corpus-based comparison of "poisonous" and "venomous." *World Journal of English Language, 15*(2), 98-115. https://doi.org/10.5430/wjel.v15n2p98

Pinner, R. (2022). Authenticity and motivation in language education: Exploring literary texts in CLT classrooms. *ELT Journal, 76*(3), 327-338. https://doi.org/10.1093/elt/ccac004

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24,* 146-161. https://doi.org/10.1017/S0267190504000078

Rosenblatt, L. (1994). *The reader, the text, the poem: The transactional theory of the literary work*. Southern Illinois University Press.

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning.* Basingstoke: Palgrave Macmillan.

Santosa, N., Pratiwi, D., & Widyantari, R. (2023). Vocabulary learning strategies and vocabulary mastery among EFL learners. *World Journal of English Language, 13*(4), 155-168. https://doi.org/10.5430/wjel.v13n4p155

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64*(4), 913-951. https://doi.org/10.1111/lang.12077

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Evidence from L2 collocation learning. *Studies in Second Language Acquisition, 35*(1), 31-65. https://doi.org/10.1017/S0272263112000678

Swain, M., & Lapkin, S. (2021). Output hypothesis revisited: Collaborative dialogue and meaning-making. *The Modern Language Journal, 105*(S1), 36-50. https://doi.org/10.1111/modl.12705

Teng, F. (2021). Incidental vocabulary learning from reading: A meta-analysis. *TESOL Quarterly, 55*(3), 910-942. https://doi.org/10.1002/tesq.3005

Tsang, A. (2023). The language and non-language benefits of literature in language education. *Language Teaching Research.* Advance online publication. https://doi.org/10.1177/13621688231163389

Uchihara, Y., & Clenton, J. (2020). Investigating the role of vocabulary depth in second language reading comprehension. *Studies in Second Language Acquisition, 42*(4), 891-915. https://doi.org/10.1017/S0272263119000614

van den Broek, P., Bohn-Gettler, C., Kendeou, P., Carlson, S., & White, M. J. (2021). When a reader meets a text: The role of inferencing in learning and memory. *Educational Psychologist, 56*(4), 245-263. https://doi.org/10.1080/00461520.2021.1923948

Webb, S., & Nation, I. S. P. (2017). How vocabulary is learned: Insights from listening input and classroom activities. *Language Teaching Research, 21*(4), 453-471. https://doi.org/10.1177/1362168816683563