

Grammar Consciousness-Raising Instruction in L2 English: A Systematic Review and Meta-Analysis

Amal Albureikan

Assistant Professor, Department of English Language and Literature, College of Languages and Humanities, Qassim University, Buraydah, Saudi Arabia

Corresponding Author: Amal Albureikan, **E-mail:** briekan@qu.edu.sa

ABSTRACT

Grammar consciousness-raising (CR) asks learners to examine linguistic data, detect patterns, and formulate or test explicit generalizations rather than receiving rules only through teacher explanation. Although CR has been investigated across diverse English-as-a-second- or foreign-language settings, its effects have not previously been isolated in a focused quantitative synthesis. This study aims to assess the effectiveness of grammar CR for improving L2 English learners' grammar performance. A systematic review and meta-analysis were conducted, drawing on 22 empirical studies involving 1,473 learners, identified through database searching (Academic Search Ultimate, Education Source, ERIC, ProQuest, Web of Science Core Collection, Wiley Online Library, and ScienceDirect), a manual literature search, and forward citation searching, from database inception to 7 August 2026. Eligible studies compared a grammar-CR treatment against a non-CR comparison condition for L2 English learners and reported sufficient quantitative evidence for a grammar-related outcome. The primary synthesis pooled the 20 of these studies that reported post-test scores (1,317 learners); grammar CR had a large overall effect on learners' grammar performance relative to comparison instruction (Hedges' $g = 0.99$, 95% CI [0.78, 1.20]). Including two further eligible studies that reported gain scores rather than post-test scores did not change this conclusion ($k = 22$, $g = 1.03$, 95% CI [0.81, 1.25]). Exploratory subgroup analysis suggested a smaller, more consistent effect in Iranian settings ($g = 0.86$) than in other settings ($g = 1.10$), although region was confounded with study design and target feature. These findings suggest that grammar CR is an effective approach for developing controlled, form-focused grammar performance, although evidence for durable transfer to spontaneous communication remains limited.

KEYWORDS

consciousness-raising; grammar instruction; focus on form; meta-analysis; second language acquisition

ARTICLE INFORMATION

ACCEPTED: 30 August 2026

PUBLISHED: 27 September 2026

DOI: 10.32996/jeltal.2026.8.10.2

1. Introduction

The role of grammar instruction remains a central problem in second-language (L2) pedagogy. Communicative approaches rightly emphasize meaning, interaction, and purposeful language use, yet learners do not always notice low-salience or complex grammatical features in ordinary input. Conversely, rule explanation and decontextualized practice may produce declarative knowledge without ensuring that learners attend to the form in subsequent comprehension or production. Research on form-focused instruction thus asks how attention to grammar can be integrated with meaningful activity. The goal is cognitive and communicative engagement, not the isolated transmission of rules (Ellis, 2001, 2006; Long, 1991; Norris & Ortega, 2000; Spada & Tomita, 2010).

Grammar consciousness-raising (CR) represents one influential response to this problem. Sharwood Smith (1981) used consciousness-raising to describe pedagogical efforts that make selected features of the target language available for conscious inspection. Rutherford and Sharwood Smith (1985) subsequently situated CR within a view of instruction as an attempt to alter what becomes salient in the learner's linguistic experience, without claiming that explicit explanation directly installs an acquired rule. In classroom applications, learners typically examine examples, compare acceptable and unacceptable forms, identify regularities, articulate a rule or hypothesis, and test that account against additional data (Fotos & Ellis, 1991; Fotos, 1993).

This problem-solving orientation distinguishes CR from several neighboring practices. It is more participatory than teacher-fronted explanation because learners analyze evidence and verbalize relationships among form, meaning, and use. It is also more explicit than incidental focus on form because the grammatical feature becomes the intentional object of inquiry. At the same time, CR need not require learners to produce the target structure immediately or accurately during the task. Its immediate pedagogical aim is understanding and noticing; productive control may develop later, through subsequent practice and exposure (Doughty & Williams, 1998; Ellis, 2002; Fotos, 1993; Swain, 1998).

The rationale for CR is commonly connected to Schmidt's (1990) account of noticing. From this perspective, linguistic input cannot contribute fully to learning unless relevant features are registered with attention. Research has subsequently examined how attention, awareness, interaction, and feedback relate to development, while also showing that noticing is difficult to observe and operationalize directly (Leow, 2000; Mackey, 2006; Robinson, 1995). CR may create a cognitive representation that helps learners recognize a target form when it reappears. However, noticing, explicit knowledge, controlled performance, and spontaneous use are not interchangeable constructs. A learner may explain a rule or succeed on an untimed grammar test while remaining unable to use the same feature rapidly in communication. Consequently, evidence for CR must be interpreted in relation to the outcome used and the distance between that outcome and the instructional task.

The broader instructional-synthesis literature provides grounds for expecting positive effects but does not answer the CR-specific question. Norris and Ortega (2000) found substantial benefits for focused L2 instruction, while also showing that effect estimates depend on research design and measurement. Spada and Tomita (2010) reported benefits of explicit instruction across both simple and complex grammatical features. Meta-analytic evidence also indicates that corrective feedback can facilitate L2 development, particularly when feedback is explicit, sustained, and measured with outcomes sensitive to the treatment (Li, 2010). These syntheses combine explanation, feedback, input processing, and production practice, so their conclusions cannot be assumed to represent the distinctive contribution of learner-led grammatical analysis.

The primary CR literature is dispersed across countries, proficiency levels, target structures, and outcome types. Studies have examined interactive grammar tasks (Fotos & Ellis, 1991), later noticing after CR treatments (Fotos, 1993), and tense and syntactic learning (Abdalla, 2014; Amirian & Sadeghi, 2012). Others have looked at writing and reading outcomes (Nosratinia & Roustayi, 2014; Qi, 1994), narrative tense performance (Tilahun et al., 2022), oral production (Dang & Nguyen, 2013), and technology-mediated collaborative writing (Khezrlou, 2024). Reported effects vary widely, and many studies use small intact classes or measures closely aligned with the treatment. A focused synthesis is therefore necessary to estimate the average effect while distinguishing grammar CR from lexical awareness, pragmatics, pronunciation, and other uses of the label consciousness-raising. The present review addresses that need by applying an explicit operational definition of grammar CR, separating strict morphosyntactic outcomes from adjacent domains, and synthesizing one independent comparison per study. In addition to estimating the pooled effect, it evaluates heterogeneity, influence, contextual patterns, methodological limitations, and the extent to which the evidence supports claims beyond immediate controlled performance.

2. Literature Review

2.1 Defining Grammar Consciousness-Raising

Consistent with established descriptions of CR (Ellis, 2002; Fotos & Ellis, 1991; Sharwood Smith, 1981), the present review classified CR according to what learners were required to do rather than relying solely on the terminology used by study authors. A prototypical CR activity presents a structured set of linguistic data and directs attention to a target relationship. Learners are then required to identify or compare patterns, elicit an explicit generalization, and check the proposed rule against further evidence. The activity may be inductive, with learners deriving the rule, or partly deductive, with a provisional rule supplied for evaluation. In both cases, active analysis is essential (Fotos & Ellis, 1991; Sharwood Smith, 1981).

Drawing this boundary matters because consciousness-raising has also been used broadly for input enhancement, awareness campaigns, pragmatic reflection, pronunciation work, corpus consultation, autonomous exercises, and teacher explanation. Such activities may be pedagogically valuable, but combining them indiscriminately would make the pooled construct uninterpretable. The present review accordingly required learner engagement with grammatical data and rule-related reasoning; exposure, highlighting, feedback, repetition, or explanation alone did not qualify as grammar CR.

2.2 Attention, Noticing, and Explicit Knowledge

The theoretical claim underlying CR is indirect. Instruction does not guarantee acquisition simply because a rule has been understood. Instead, explicit analysis may orient attention, increase the perceptual salience of a form, and provide a hypothesis that learners can test against later input. Schmidt (1990) argued that noticing is a necessary condition for converting aspects of input into intake, and Fotos (1993) found that both grammar problem-solving tasks and formal instruction increased subsequent noticing relative to a control condition. CR is thus plausibly understood as preparation for further processing rather than as a complete learning sequence in itself.

A second distinction concerns explicit and implicit knowledge. The two knowledge systems are theoretically distinguishable but may interact during learning and performance (N. C. Ellis, 2005). CR directly targets conscious understanding and metalinguistic awareness, so its most proximal outcomes are rule verbalization, grammaticality judgment under low time pressure, and controlled use of a target structure. Test batteries designed to separate the constructs associate untimed judgment and metalinguistic explanation more strongly with explicit knowledge, whereas oral narrative, elicited imitation, and timed judgment

place greater demands on implicit knowledge (R. Ellis, 2005; Zhang, 2015). Stronger evidence of proceduralization therefore requires delayed measurement, time pressure, novel items, or spontaneous production. Without such measures, a positive post-test should be described as improved form-focused performance rather than proof that the target feature has become automatic.

2.3 CR as a Communicative Grammar Task

CR does not necessarily conflict with communicative pedagogy. Fotos and Ellis (1991) demonstrated that learners could negotiate a difficult grammatical problem while using the L2 for interaction, allowing grammar to become the content of communication. The pedagogical value lies in a dual focus: learners exchange meanings and manage interaction while collaboratively constructing knowledge about form. That dual focus is consistent with proposals to integrate brief, responsive attention to form into meaning-oriented lessons (Long, 1991) and with evidence that learner uptake and learning opportunities can follow incidental form-focused episodes (Loewen, 2005). Such tasks contrast with exercises in which learners merely manipulate predetermined sentences after a rule has been supplied.

Nevertheless, communicative interaction during a CR task should not be confused with communicative mastery of the target feature. Learners may communicate successfully about a rule without using that rule accurately in unplanned speech. An adequate evidence base must therefore distinguish the interaction generated by the learning task from later transfer of the target form to new communicative contexts.

2.4 Empirical Evidence and Sources of Variation

Individual CR studies generally report positive results, but the size and meaning of those results differ. Controlled investigations have targeted a range of linguistic features, including tense and aspect, syntax, the copula, and morphological awareness (Abdalla, 2014; Amirian & Sadeghi, 2012; Heidari & Mansourzadeh, 2014; Idek et al., 2014). Others have examined analytic writing, narrative writing, and oral grammar (Tashmuradova et al., 2023; Tilahun et al., 2022). Other work has embedded CR in collaborative or technology-mediated writing and has attempted to distinguish explicit from implicit outcomes (Khezrlou, 2024). This diversity suggests that treatment effects may depend on the linguistic feature, the comparison condition, treatment duration, alignment between instruction and test, and the degree of productive demand.

Context and design may also shape estimates. Much of the available evidence comes from EFL university settings and intact classes. Small samples increase uncertainty, while class-level assignment can yield overly precise estimates when clustering is ignored. Studies also differ in whether comparison groups receive conventional instruction, alternative form-focused treatment, or minimal intervention. Research comparing prompts, recasts, and metalinguistic feedback further demonstrates that instructional contrasts are not interchangeable (Ellis et al., 2006; Lyster, 2004). Finally, immediate target-specific tests are likely to show larger effects than delayed, untrained, or spontaneous measures. These features make a random-effects synthesis appropriate and require caution when interpreting moderators.

2.5 Need for the Present Review

Existing meta-analyses establish that explicit and form-focused L2 instruction is effective in general (Norris & Ortega, 2000; Spada & Tomita, 2010), but they do not isolate grammar CR as an analytically distinct intervention. A CR-specific synthesis must apply a defensible definition, prevent multiple outcomes from the same study from being treated as independent, and examine whether positive findings are robust to influential studies. It must also avoid extending evidence from controlled grammar tests to implicit competence or spontaneous communication without appropriate measures. The present systematic review and meta-analysis was designed around these requirements.

The review addressed three questions:

1. What is the pooled effect of grammar CR instruction on L2 English outcomes relative to comparison instruction?
2. How heterogeneous is the effect, and what contextual patterns are visible?
3. What features of the evidence base support or limit claims about retention and communicative transfer?

3. Methodology

3.1 Searching Process

The review was reported with reference to PRISMA 2020 (Page et al., 2021). All eligibility criteria, search procedures, extraction decisions, and analytical methods are reported to support transparency and reproducibility. The review questions, operational definition, extraction fields, and effect-selection rules were fixed before the final synthesis.

Searches were completed on 7 August 2026 from database inception in Academic Search Ultimate, Education Source, and ERIC through EBSCO; ProQuest; Web of Science Core Collection; and Wiley Online Library. ScienceDirect was searched as a supplementary publisher platform. Search terms combined CR expressions with English-learning and controlled-design terms. A representative syntax was: ("consciousness-raising" OR "consciousness raising" OR "grammar consciousness-raising" OR "consciousness-raising task*") AND (English OR EFL OR ESL OR "L2 English") AND (experiment* OR intervention OR pretest OR posttest OR "control group" OR "comparison group"). Reference lists of eligible reports were also checked.

3.2 Inclusion and Exclusion Criteria

Eligibility was defined at two levels. Studies meeting the population, intervention, comparator, and design criteria were retained in the systematic review. Inclusion in the primary meta-analysis additionally required a morphosyntactic English outcome and

sufficient statistics for an independent standardized mean difference. Adjacent studies of lexical or pragmatic consciousness-raising were retained only to map the boundary of the literature and were not pooled with grammar effects.

Table 1. Inclusion and exclusion criteria for the systematic review and meta-analysis.

Domain	Included	Excluded
Population	Learners of English as a second, foreign, or additional language at any proficiency or educational level.	First-language English populations; teacher-only samples; studies not involving English learning.
Intervention	Grammar CR requiring conscious attention to linguistic data plus learner comparison, pattern detection, hypothesis formation, rule articulation, or rule testing.	Rule explanation alone; input enhancement alone; corrective feedback, repetition, autonomous practice, or general form-focused instruction without learner analysis.
Comparator	A concurrent non-CR control or comparison condition, including conventional grammar teaching, alternative instruction, or no-treatment/business-as-usual control.	Single-group pretest-post-test designs; comparisons between two CR variants without a non-CR reference condition.
Outcome	Quantitative L2 English morphosyntactic learning outcome, such as grammar accuracy, grammaticality judgment, morphology, writing grammar, or oral grammatical production.	Vocabulary/collocation, pragmatics, pronunciation, attitudes, awareness only, or interaction counts without a grammar-learning outcome. Adjacent outcomes could enter narrative review only.
Design	Controlled experimental or quasi-experimental primary study with post-intervention group data or convertible statistics.	Qualitative studies, case reports, reviews, theoretical papers, protocols, commentaries, and studies without extractable comparative data.
Report	Journal article, dissertation/thesis, or retrievable scholarly report in English; no date restriction.	Duplicate reports of the same sample; inaccessible reports with insufficient bibliographic or outcome information.
Meta-analysis rule	Sufficient sample sizes, means and dispersion statistics, adjusted estimates, or another statistic convertible to Hedges' <i>g</i> ; one independent effect per study.	Insufficient quantitative information after checking the report; statistically dependent effects that could not be combined or selected using the prespecified hierarchy.

3.3 Screening Process

Titles and abstracts were screened against the eligibility criteria in two stages. The database search was supplemented with a manual literature search and forward citation searching, following PRISMA 2020 guidance for reviews that combine database and other-methods identification streams (Page et al., 2021). The database search generated 526 field-level database hits (Academic Search Ultimate, *n* = 40; Education Source, *n* = 51; ERIC, *n* = 74; ProQuest, *n* = 238; Web of Science Core Collection, *n* = 42; Wiley Online Library, *n* = 73; ScienceDirect, supplementary, *n* = 8). Of these, 368 unique exportable records remained after duplicate and internally repeated hits were removed. An initial title-and-abstract screen excluded 283 clearly ineligible records, leaving 85 candidate records for detailed title-and-abstract assessment. That detailed assessment excluded a further 67 records, and 18 reports proceeded to full-text assessment; one full-text report was subsequently excluded for not meeting the eligibility criteria, leaving 17 reports retained from the database stream. All 85 candidate records that underwent detailed title-and-abstract assessment were independently screened by two reviewers. They agreed on 80 decisions (94.1% raw agreement), and Cohen's kappa indicated strong inter-rater reliability ($\kappa = .83$). The five disagreements were resolved through discussion. A separate, non-database identification stream combined a manual/supplementary literature search with forward citation searching; Appendix B gives the full procedure and record lists. The manual/supplementary search directly contributed 10 of the 28 reports that were, at that point, eligible or candidate seeds for citation-chasing. Forward-searching the citation graphs of all 28 seeds on Google Scholar, conducted on 9 August 2026, then surfaced a further 18 records (17 new candidates plus 1 duplicate of an already-included report, removed on identification). Combining these two sources, the other-methods stream identified 28 records in total, of which 27 were unique after deduplication and sought for retrieval. Eight could not be retrieved despite reasonable effort (Appendix B.4), leaving 19 reports assessed for eligibility, of which 7 were excluded (Appendix B.5) and 12 retained. Twenty-six of the 28 seed reports were themselves successfully forward-searched; the reasons the remaining two could not be searched, and a third was temporarily blocked, are given in Appendix B.3. Combining both identification streams, 29 reports were retained in the review overall.

At title-and-abstract screening, records were excluded when they clearly did not concern L2 English learning, did not implement grammar consciousness-raising, lacked an empirical comparison, or addressed an unrelated disciplinary use of consciousness-raising. Full texts were then assessed against every criterion in Table 1. Reasons for exclusion were recorded in the screening log, and duplicate reports of the same sample were linked so that a sample could contribute only once to the quantitative synthesis.

3.4 Document Coding

For each study, the review extracted country, target feature, intervention and comparison conditions, design, sample sizes, outcome timing, means, standard deviations, and convertible statistics. If a study contained more than one eligible CR arm, eligible arms were combined before comparison with a shared control. The primary outcome hierarchy was: adjusted endpoint; immediate post-test controlled grammar score; latest immediate target-specific score; and gain score when endpoint dispersion was unavailable. One independent effect was retained per study.

Coding also distinguished the variables used for descriptive and exploratory analyses: publication year, country or region, educational setting, learner level when reported, grammatical target, delivery format, treatment duration, assessment construct, assessment timing, and effect-score basis. A source-status field documented whether values came directly from an original table, were reconstructed from another reported statistic, or required a prespecified selection decision. The complete coded dataset accompanies the manuscript in spreadsheet form.

3.5 Methodological Appraisal

The MMAT 2018 category corresponding to each study design was applied criterion by criterion (Hong et al., 2018). Results were expressed as Yes, No, or Can't tell, without a summed quality score. Appraisal focused on representativeness, measurement, completeness, confounding or baseline comparability, and fidelity to the assigned intervention. Because many studies used intact classes without intraclass correlations, cluster adjustment could not be implemented consistently. In addition to this general quality screen, each study was assigned a structured risk-of-bias judgment using RoB 2 (Sterne et al., 2019) for studies with reported random assignment and ROBINS-I (Sterne et al., 2016) for non-randomized studies, following each tool's domain structure and standard algorithm for reaching an overall judgment (Appendix A.1).

4. Data Analysis

4.1 Effect-Size Calculation and Statistical Model

Standardized mean differences were corrected to Hedges' g , with positive values favoring CR (Borenstein et al., 2009; Hedges, 1981). Sampling variances used the conventional independent-groups approximation. Effects were pooled using a restricted maximum-likelihood (REML) random-effects model. Because the number of studies was small, Hartung-Knapp t -based confidence intervals were used, computed directly from the REML τ^2 estimate following the standard Hartung-Knapp-Sidik-Jonkman formula (Int'Hout et al., 2014). Heterogeneity was reported as τ^2 , τ , I^2 (Higgins et al., 2003), and a 95% prediction interval, which describes the range of effects expected across comparable settings rather than only uncertainty around the mean (Riley et al., 2011). Leave-one-study-out analyses, a funnel plot, and an exploratory regression test for funnel asymmetry (Egger et al., 1997) were calculated. Region subgroups were exploratory and were not treated as causal moderators.

4.2 Heterogeneity and Contextual Analyses

Cochran-type dispersion was summarized with I^2 , while τ^2 and τ represented the estimated between-study variance and its standard-deviation scale. A 95% prediction interval was calculated to show the range in which a true effect from a comparable future setting would be expected. Because the number of studies was limited and study characteristics overlapped, subgroup estimates by region were treated as descriptive and hypothesis-generating rather than as evidence of causal moderation.

4.3 Publication Bias and Sensitivity Analysis

Potential small-study and reporting effects were explored using a funnel plot and a regression-based asymmetry test. These diagnostics were interpreted cautiously because tests of funnel asymmetry have low power with a small corpus and cannot distinguish publication bias from genuine heterogeneity. Robustness was examined through leave-one-study-out analysis, recalculating the pooled effect after removing each study in turn. No trim-and-fill correction was treated as a definitive bias adjustment because its assumptions are difficult to defend in a heterogeneous instructional literature.

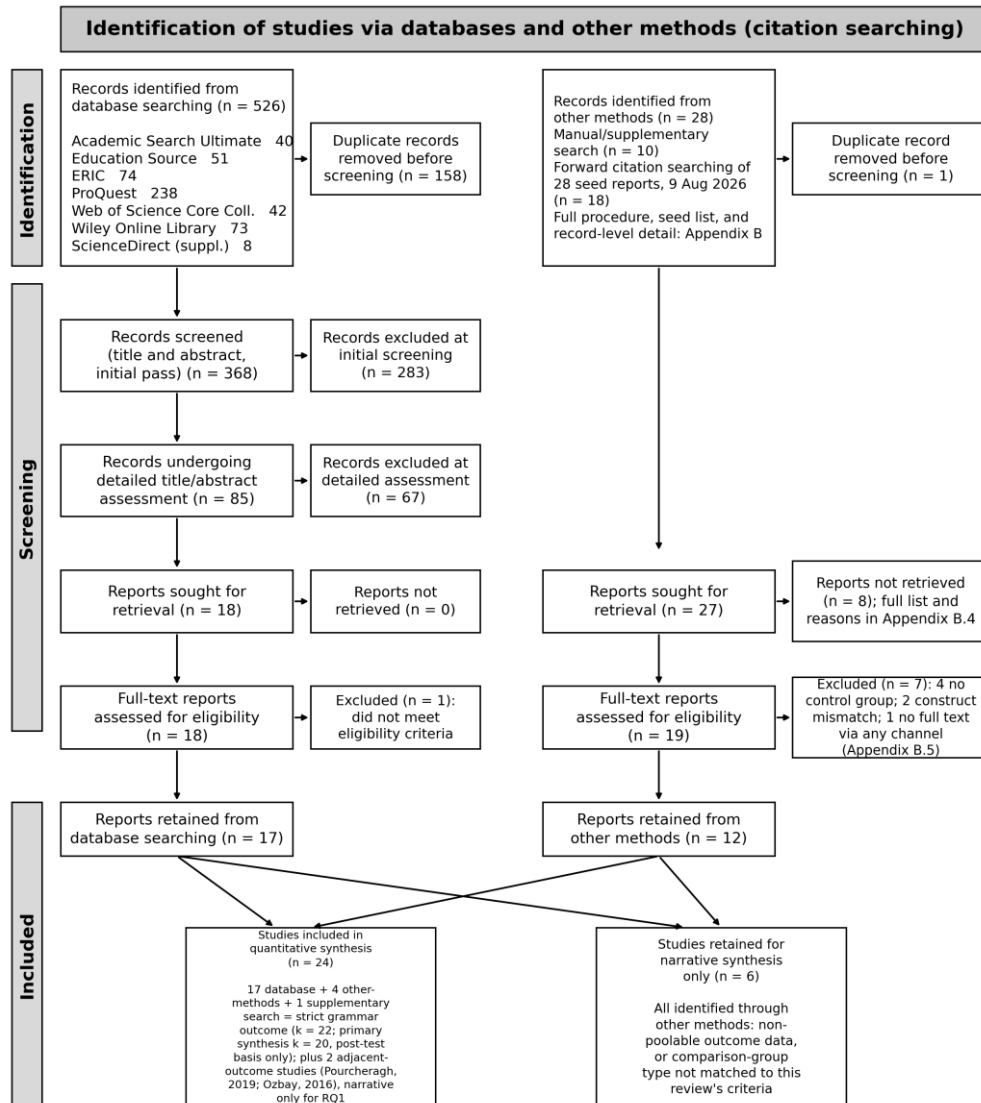
5. Findings

5.1 Study Selection

Twenty-two studies published between 1994 and 2025 met the strict grammar-outcome eligibility criteria ($N = 1,473$). The primary synthesis pooled the 20 of these studies that reported post-test scores ($N = 1,317$); the remaining two, which reported gain scores rather than post-test scores, were included in a separate sensitivity analysis (Section 5.5) rather than mixed with the post-test-only studies. Nine were conducted in Iran; the remaining thirteen studies were conducted in Afghanistan, Ethiopia, Indonesia, Japan, Malaysia, Oman, Saudi Arabia, South Korea, Turkey, the United States, and Vietnam. Outcomes included controlled grammar tests, grammaticality judgments, morphological awareness, analytic writing, narrative writing, oral grammatical production, article and preposition tests, and writing accuracy.

Six additional controlled studies met the general grammar-CR eligibility criteria but did not supply a single, non-arbitrary, extractable between-group post-test effect for pooling, and are therefore discussed narratively rather than pooled. Azizifar, Babaei, Jamalinesari, and Gowhary (2015), Xavier and Gesser (2022), and Hosseininik and Sangani (2014) reported genuine CR-versus-comparison designs but presented divergent or non-independent outcome cells, or within-group rather than between-group statistics only. No defensible single effect could therefore be selected from these three studies without risking selective-outcome-reporting bias. Erfani and Torkamani (2015), identified through forward citation searching, compared CR against an alternative structure-based production task rather than a traditional or no-treatment control on a reading-comprehension outcome; CR did not outperform the comparison group on that outcome. Fazlali and Shahini (2019) and Namaghi and Charmchi

(2016), also identified through forward citation searching, reported a collocation-knowledge outcome and an omnibus multi-group ANOVA without per-group post-test descriptives, respectively, neither of which is extractable as this review's morphosyntactic single-effect outcome. Two further studies contributed extractable quantitative data but were retained outside the primary strict-grammar synthesis on outcome-type grounds. Pourcheragh (2019) compared consciousness-raising instruction against a comparison condition for Iranian EFL upper-intermediate learners' pragmatic development of agreeing and disagreeing in speaking performance ($n = 30$ per arm; Hedges' $g = 0.05$), an oral-pragmatic rather than a morphosyntactic outcome. Özbay (2016) compared corpus-based explicit collocation instruction against a comparison condition for Turkish EFL learners' collocation awareness and reading performance ($n = 25$ per arm; Hedges' $g = 2.48$), a lexical-collocational and reading outcome rather than a morphosyntactic one. Both were included in an all-outcomes sensitivity synthesis ($k = 24$, $g = 1.04$, 95% CI [0.80, 1.29]) reported alongside the primary strict-grammar estimate in Section 5.3. They were excluded from the primary synthesis on the operational-definition boundary set out in Section 2.1, which separates grammar CR from adjacent domains such as pragmatics and lexis.



Note: a supplementary targeted search on 10 August 2026 (Appendix B.6), conducted in response to peer-review feedback, identified 2 further records outside the flow above: 1 included (reflected in the totals shown here) and 1 excluded for construct mismatch (Moradkhani & Sohrabian, 2009). Two of the 22 strict grammar-outcome studies reported gain scores rather than post-test scores; the primary synthesis pools the 20 post-test-only studies, with the full 22-study set reported as a sensitivity analysis (Section V.E).

Figure 1. PRISMA 2020 study-selection flow diagram.

5.2 Characteristics of the Included Studies

Table 2. Characteristics and primary effect estimates of included studies. The two studies whose Basis is “Gain score” (Idek et al., 2014; Abdalla, 2014) are reported for completeness but were not pooled in the primary post-test-only synthesis (Section 5.2); they are included only in the $k = 22$ sensitivity analysis (Section 5.5).

Study	Country	Primary outcome	NCR/control	Basis	g [95% CI]
Amirian & Sadeghi (2012)	Iran	Controlled grammar test	30/30	Post-test	0.61 [0.09, 1.13]
Heidari & Mansourzadeh (2014)	Iran	Controlled grammar test	51/54	Post-test	0.55 [0.16, 0.94]
Behrouzi & Kazemirad (2012)	Iran	Controlled grammar/syntax test	30/30	Post-test	0.87 [0.34, 1.40]
Idek et al. (2014)	Malaysia	Verb be grammar test	28/28	Gain score	0.68 [0.14, 1.22]
Abdalla (2014)	Saudi Arabia	Tense test	50/50	Gain score	1.99 [1.51, 2.47]
Fatemipour & Hemmati (2015)	Iran	Controlled grammar test	30/30	Post-test	1.12 [0.57, 1.67]
Tilahun et al. (2022)	Ethiopia	Narrative-tenses writing	45/45	Post-test	1.30 [0.84, 1.76]
Dang & Nguyen (2013)	Vietnam	Oral grammatical production	47/47	Post-test	0.57 [0.16, 0.98]
Tashmuradova et al. (2023)	Afghanistan	Morphological awareness	30/30	Adjusted post-test estimates	1.64 [1.05, 2.22]
Nosratinia & Roustayi (2014)	Iran	Analytic writing ability	30/30	Post-test	1.26 [0.71, 1.82]
Qi (1994)	United States	Sentence-structure grammaticality judgment	9/22	Post-test	0.70 [-0.10, 1.50]
Yarahmadzahi et al. (2015)	Iran	Controlled grammar proficiency	42/24	Post-test	0.62 [0.11, 1.13]
Khezrlou (2024)	Iran	Untimed grammaticality judgment	36/24	Immediate post-test	1.58 [0.98, 2.17]
Ma & Cho (2019)	South Korea	Controlled grammar test	51/26	Immediate post-test	1.06 [0.56, 1.57]
Hassanzadeh & Shahbazi (2021)	Iran	Explicit article recognition/production	30/25	Immediate post-test	0.34 [-0.19, 0.86]
Öncü (1998)	Turkey	Modals (can/may/must/have to)	30/30	Post-test (t-reconstructed)	2.00 [1.39, 2.62]
Al-Lawatiyah (2012)	Oman	Prepositions of time and paired adjectives	22/18	Post-test	1.37 [0.69, 2.05]
Brender (2002)	Japan	English articles, cloze test	56/35	Post-test	0.63 [0.20, 1.06]
Pakbaz & Rezai (2015)	Iran	Writing accuracy (T-unit based)	31/29	Immediate post-test	1.02 [0.49, 1.55]
Rahayu (2018)	Indonesia	Grammar mastery (multiple-choice)	17/17	Post-test	0.86 [0.17, 1.54]
Iskandar & Heriyawati (2015)	Indonesia	Grammatical competence (present perfect)	35/35	Post-test	0.58 [0.10, 1.05]
Al-Mubireek (2025)	Saudi Arabia	Grammatical accuracy (English tenses)	43/41	Post-test	1.56 [1.07, 2.04]

5.3 Overall Effect and Heterogeneity

Grammar CR produced a large average advantage over comparison instruction in the primary post-test-only synthesis ($k = 20$), $g = 0.99$, 95% CI [0.78, 1.20]. Between-study variance was $\tau^2 = 0.13$ ($\tau = 0.36$), and $I^2 = 64.3\%$, indicating substantial heterogeneity. The 95% prediction interval was [0.21, 1.78]. Thus, the estimated mean and the predicted range of true effects were positive, but the magnitude expected in a new setting remains uncertain. Two further eligible studies (Idek et al., 2014; Abdalla, 2014) reported gain scores rather than post-test scores; because gain scores and post-test scores are not always directly comparable in magnitude, these two studies were not pooled with the primary post-test-only synthesis. As a sensitivity check, adding them to the primary 20 studies produced a $k = 22$ estimate of $g = 1.03$, 95% CI [0.81, 1.25], $\tau^2 = 0.17$, $I^2 = 70.6\%$, indicating that the direction and approximate magnitude of the effect did not depend on this metric choice. As a further sensitivity check, adding the two adjacent-outcome studies excluded from the strict synthesis (Pourcheragh, 2019; Özbay, 2016) to the full 22-study set produced an all-outcomes estimate of $k = 24$, $g = 1.04$, 95% CI [0.80, 1.29], $\tau^2 = 0.25$, $I^2 = 77.3\%$. This confirms that the primary conclusion is not sensitive to the gain-score/post-test boundary or the adjacent-outcome boundary.

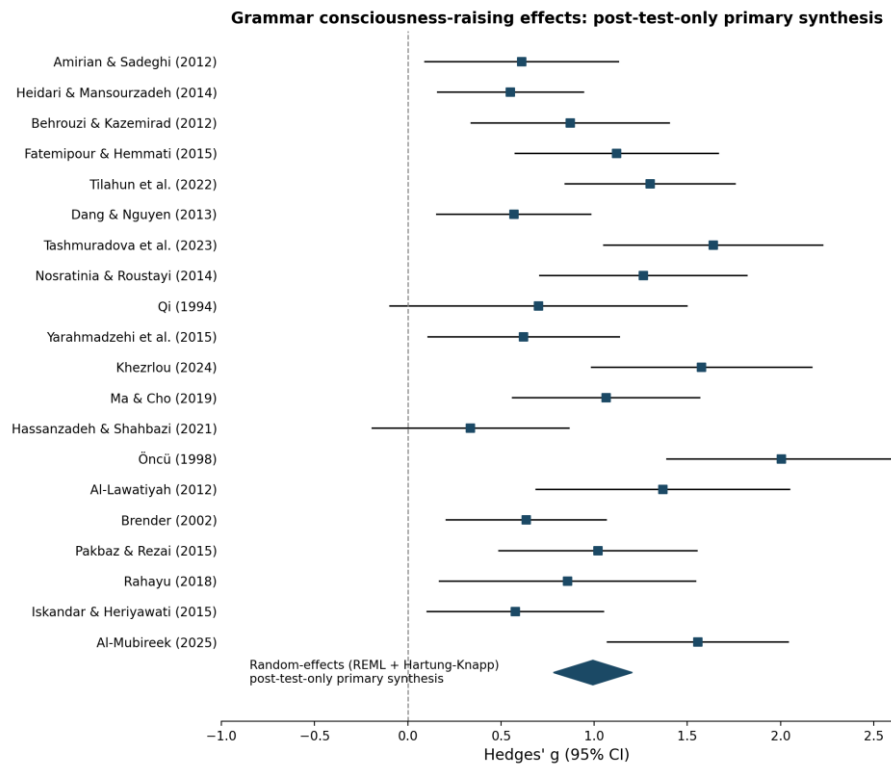


Figure 2. Forest plot of the primary post-test-only synthesis ($k = 20$; REML random effects with Hartung-Knapp interval).

5.4 Exploratory Contextual Patterns

Within the primary post-test-only synthesis, the nine Iranian studies yielded $g = 0.86$, 95% CI [0.56, 1.16], $I^2 = 52.2\%$. The remaining eleven studies from other settings yielded $g = 1.10$, 95% CI [0.77, 1.43], $I^2 = 69.7\%$. These estimates should not be interpreted as evidence of a regional cause because region is confounded with design, outcome, sample, and target structure, and each subgroup remains small.

5.5 Publication-Bias and Influence Diagnostics

Because many included studies used intact classes without reporting intraclass correlations (Section 5.6), clustering that was not modeled in the primary variance estimates could in principle have made the confidence intervals appear narrower than they should be. As a further sensitivity check, each study's variance was inflated using a design-effect correction, $DEFF = 1 + (m - 1) \times ICC$, with m approximated by the average arm size and ICC values of 0.05, 0.10, and 0.20, spanning the typical range reported for education achievement outcomes (Hedges & Hedberg, 2007). Under these assumptions, the primary pooled estimate changed only marginally, $g = 0.97$ across all three ICC values (95% CIs of approximately [0.76, 1.18], overlapping the unadjusted estimate), because the higher assumed within-study variance was offset by a corresponding reduction in the estimated between-study variance. This indicates that the primary conclusion is unlikely to be an artifact of unmodeled clustering, although the true design effect for each study remains unknown because none of the included reports specified the number of clusters or an empirical intraclass correlation.

Removing one study at a time did not change the positive direction of the pooled effect. Omitting any single study from the primary 20-study synthesis, the pooled estimate ranged from $g = 0.94$ (without Öncü, 1998) to $g = 1.03$ (without Hassanzadeh & Shahbazi, 2021), and every leave-one-study-out confidence interval excluded zero. The result was thus not dependent on a single report, although omission of studies with very large effects reduced the estimated magnitude somewhat. The regression test for funnel asymmetry approached but did not reach conventional significance in the primary post-test-only synthesis (intercept = 4.24, $t = 2.02$, $p = .058$); in the $k = 22$ sensitivity synthesis that also included the two gain-score studies, the same test was clearly non-significant (intercept = 3.62, $t = 1.51$, $p = .148$). Both tests have low power with a corpus of this size and do not rule out selective reporting; the borderline result in the primary synthesis is noted here as a caution rather than a confirmed asymmetry.

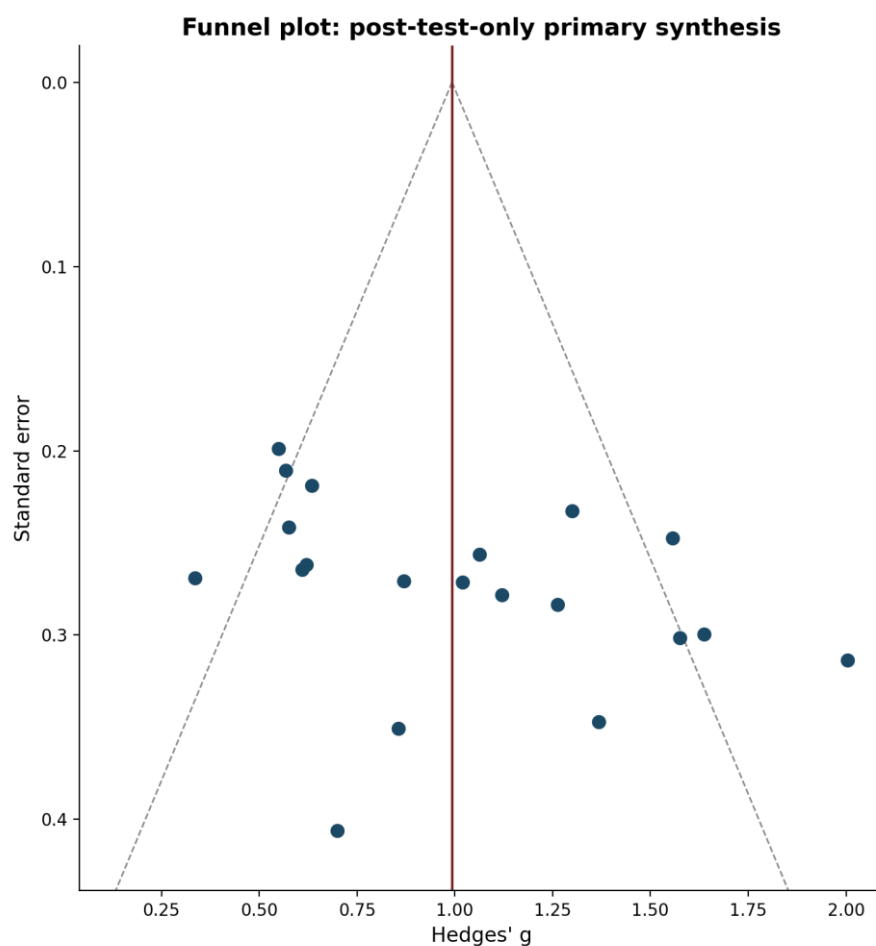


Figure 3. Funnel plot for the primary post-test-only synthesis (k = 20).

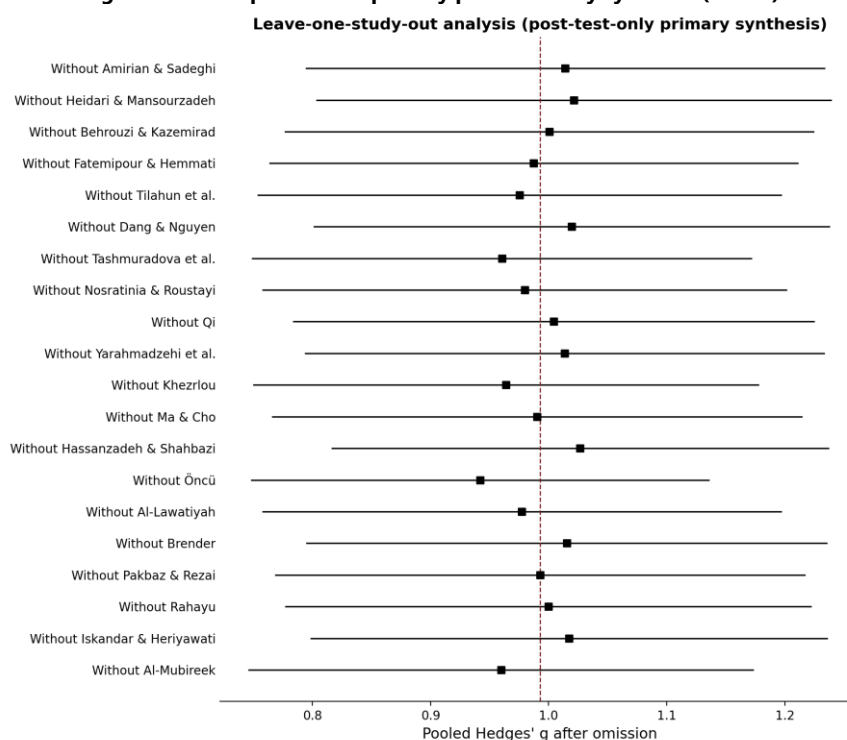


Figure 4. Leave-one-study-out sensitivity analysis (primary post-test-only synthesis, k = 20).

5.6 Methodological Appraisal Findings

The evidence base was methodologically uneven. Most of the 22 studies used small samples and intact classes, and many did not report allocation procedures, assessor blinding, intraclass correlations, or complete instrument properties. Baseline equivalence was established in some reports but unclear in others. The largest effects often came from outcomes closely aligned with the trained material or from comparisons in which the control group changed little. MMAT results were therefore interpreted by criterion and not converted to numerical quality scores. A full criterion-level appraisal for each study is provided in Appendix A. The RoB 2 / ROBINS-I appraisal (Appendix A.1, Tables 4a-4b) produced a more decisive picture: both randomized studies were judged at an overall rating of some concerns, and 18 of the 20 non-randomized studies were judged at serious overall risk of bias, driven mainly by uncontrolled confounding in the absence of random allocation. Only two non-randomized studies, which reported evidence of baseline comparability between groups (a direct pre-test equivalence test in one case, and no significant pre-test difference on the target measure in the other), were judged at moderate rather than serious risk. This pattern indicates that the pooled effect should be read as an estimate from a body of evidence with substantial, design-driven risk of bias, rather than as a precise causal parameter.

6. Discussion

6.1 Principal Findings

The strict synthesis provides a clear but bounded answer to the first research question. Across the 20 controlled studies pooled in the primary post-test-only synthesis (1,317 learners), grammar consciousness-raising produced a large average advantage over comparison instruction, $g = 0.99$, 95% CI [0.78, 1.20]. This conclusion held when two further studies that reported gain scores were added as a sensitivity check ($k = 22$, $g = 1.03$). The confidence interval excluded zero, the prediction interval remained positive, and every leave-one-study-out estimate favored CR. These converging results indicate that the pooled effect was not created by a single unusually favorable report. They support the proposition that requiring learners to inspect linguistic evidence, compare alternatives, articulate a generalization, and test that generalization can improve performance on grammar-related outcomes.

The magnitude should nevertheless be interpreted as an average across a heterogeneous and methodologically uneven literature, rather than as a universal treatment constant. An effect of $g = 0.99$ does not imply that every learner, grammatical feature, or instructional setting will obtain an improvement of that size. The prediction interval [0.21, 1.78] is especially informative: it suggests that a comparable future study would most plausibly produce a positive effect, while also allowing a wide range from small to very large benefits. The substantive conclusion is therefore stronger with respect to the direction of the effect than to its precise magnitude.

6.2 Relationship to Previous Instructional Research

The present result is broadly consistent with earlier syntheses showing that focused and explicit L2 instruction is effective (Norris & Ortega, 2000; Spada & Tomita, 2010). It also accords with research demonstrating that pedagogical interventions that make form salient, through explanation, feedback, prompts, or structured attention, can facilitate development (Ellis, 2001, 2006; Li, 2010; Lyster, 2004). The contribution of the present synthesis is narrower: it isolates interventions in which learners actively analyze grammatical data rather than treating all explicit or form-focused procedures as equivalent. The positive pooled estimate accordingly supports CR as a distinct instructional option within the broader family of form-focused approaches.

The distinction matters because CR is not simply another label for rule explanation. In prototypical CR, the learner performs the analytic work: detecting a pattern, comparing forms, proposing a rule, and checking it against further evidence. That architecture is visible in the grammar problem-solving tradition represented by Fotos and Ellis (1991) and Fotos (1993), and it remains central to later classroom applications. The results suggest that this active analytic component can add value when compared with conventional explanation, drills, textbook-led practice, or other non-CR conditions. Because the comparison treatments varied considerably, however, the synthesis does not identify a single instructional component as the exclusive source of the advantage.

6.3 Theoretical Interpretation

The findings are compatible with noticing-based accounts, but they do not constitute a direct test of noticing as a cognitive mechanism. CR deliberately directs attention toward relationships among form, meaning, and use, which is consistent with the claim that attended features are more likely to become available for subsequent processing (Schmidt, 1990). Collaborative analysis may additionally require learners to verbalize hypotheses and notice gaps in their developing system, while repeated comparison with new examples may strengthen the representation of a target feature (Fotos, 1993; Swain, 1998). These processes offer a plausible explanation for the observed gains, but most included studies measured learning products rather than attention or awareness during the intervention.

The results are also compatible with an interface position in which explicit knowledge can support later processing without automatically becoming implicit knowledge (N. C. Ellis, 2005). CR most directly develops conscious understanding: learners become better able to recognize a pattern, explain a contrast, monitor a response, or deliberately apply a rule. Such knowledge can guide practice and may contribute to proceduralization when learners encounter the feature repeatedly under increasingly communicative and time-pressured conditions. The present evidence cannot determine whether that transition occurred because few studies combined controlled tests with delayed, timed, or spontaneous measures.

6.4 Outcome Measurement and Communicative Transfer

The meaning of the pooled effect depends heavily on what the included assessments measured. Most outcomes were immediate and allowed deliberate attention to form. Controlled grammar tests, untimed grammaticality judgments, and exercises closely aligned with the taught material are legitimate indicators of form-focused achievement, but they are particularly sensitive to explicit knowledge and strategic monitoring (R. Ellis, 2005; Zhang, 2015). High performance on such measures should not be equated automatically with rapid access to grammar during spontaneous communication. Treatment-test similarity may also enlarge effects when assessment items closely resemble the examples or structures practiced during instruction.

Some studies extended assessment beyond isolated grammar items. Dang and Nguyen examined oral grammatical production, Tilahun and colleagues assessed narrative-tense writing, Nosratinia and Roustayi included analytic writing, and Khezlou attempted to distinguish explicit and implicit outcomes. These studies broaden the evidential base, but they do not yet form a sufficiently large or methodologically uniform subgroup to establish reliable transfer. Delayed spontaneous outcomes, untrained items, time-pressured production, and ecologically valid classroom communication remained uncommon. The third research question must therefore be answered cautiously: evidence supports gains in controlled and form-focused performance, whereas durable, automatic, and communicative use remains insufficiently tested rather than disproved.

6.5 Sources and Meaning of Heterogeneity

The substantial heterogeneity addresses the second research question and is theoretically plausible. The studies differed in learner age and proficiency, grammatical target, instructional duration, individual versus collaborative work, comparison condition, and assessment format. A CR treatment compared with minimal or routine instruction does not represent the same contrast as CR compared with a well-developed alternative form-focused treatment. Similarly, a target-specific immediate test is likely to yield a different estimate from a broad writing measure or oral production task. Variation in effect sizes should therefore be understood partly as variation in the instructional and measurement contrasts being estimated, rather than as random noise alone.

The exploratory regional estimates do not justify a cultural or national explanation. Iranian and non-Iranian studies differed simultaneously in design, target feature, educational level, measurement, and effect magnitude; with only nine and twelve studies respectively, those characteristics cannot be disentangled. Region is thus a descriptive grouping variable, not a causal moderator. The same caution applies to any apparent advantage associated with a particular outcome or delivery format when categories contain few studies and overlap with study quality.

6.6 Methodological Strength of the Evidence

The MMAT appraisal qualifies the strength of causal interpretation. Only a small minority of studies documented random allocation, and even those reports frequently omitted details on sequence generation, allocation concealment, or assessor blinding. Many investigations used intact classes or groups formed after proficiency screening. Such designs can provide useful classroom evidence, but baseline differences, teacher effects, institutional context, and self-selection may contribute to post-intervention contrasts when confounders are not fully controlled. Cluster-based allocation also creates dependence among learners in the same class; because intraclass correlations were not reported, the meta-analysis could not apply a consistent design-effect correction.

Several further features may have increased apparent precision or effect magnitude: small samples, selective reporting of favorable outcomes, incomplete description of attrition, outcomes designed by the investigators, and close alignment between treatment activities and post-tests. The non-significant funnel-asymmetry test does not remove these concerns because such tests have limited power with 15 heterogeneous studies. Conversely, leave-one-study-out analysis showed that the positive conclusion did not depend on any single report. Taken together, the evidence supports a probable instructional benefit while warranting less confidence in the exact pooled magnitude and in strong claims of causality, durability, or generalizability.

6.7 Overall Interpretation

The most defensible interpretation integrates the statistical, theoretical, and methodological evidence. Grammar CR is effective as an analytic instructional phase that strengthens learners' conscious control of L2 form. Its value lies in converting grammar from information delivered by the teacher into a problem investigated by learners. However, CR should not be treated as a complete acquisition sequence or as a substitute for communicative practice. The evidence is most consistent with a staged account in which analysis and hypothesis formation establish an explicit representation that must subsequently be practiced, retrieved, and applied under progressively less controlled conditions. This interpretation preserves the substantial positive finding while remaining aligned with what the included studies actually measured.

7. Implications and Future Research

Teachers may use CR when a lesson requires learners to analyze a recurrent grammatical problem, compare examples, and articulate a generalization. The evidence does not support replacing communicative practice with analysis alone. A defensible sequence is noticing and hypothesis formation followed by structured production, communicative use, and delayed recycling.

Future trials should preregister outcomes, report group means and standard deviations at every time point, document assignment and attrition, account for class clustering, and include delayed tests and spontaneous production. Multi-arm studies should report statistics that permit dependent-effect modeling. Trials should also separate performance on trained items from

transfer to untrained contexts. A larger, prospectively registered corpus is also needed before moderator variables (grammatical target, learner level, treatment duration, delivery format) can be modeled formally rather than described narratively: with 20 studies distributed across many overlapping design features, a formal meta-regression would be underpowered and prone to overfitting. Consistent registration and reporting of publication-bias diagnostics, comparable across studies, would also let future reviews use funnel-asymmetry and selection-model methods with more confidence than the exploratory tests reported here.

8. Limitations

The search covered all prespecified information sources: Academic Search Ultimate, Education Source, ERIC, ProQuest, Web of Science Core Collection, Wiley Online Library, and a supplementary ScienceDirect search. It was conducted from database inception to 7 August 2026 and was complemented by a manual literature search and forward citation searching. The review's breadth here is a strength. Nevertheless, no search can guarantee capture of every relevant report because database coverage, indexing terminology, availability of older full texts, and publisher export functions differ. The limitation thus concerns the possibility of residual retrieval gaps, not failure to search the databases specified in the search plan.

Formal inter-rater reliability was established at the title-and-abstract screening stage and, subsequently, for the risk-of-bias appraisal: a second, independent rater re-applied RoB 2 / ROBINS-I to all 22 pooled studies' overall judgments, yielding 20/22 (90.9%) raw agreement. The two discrepancies (Behrouzi & Kazemirad, 2012; Iskandar & Heriyawati, 2015) were resolved through discussion and are reflected in the ratings reported in Table 4b and its accompanying note. Full-text eligibility decisions and data extraction were not independently duplicated, relying instead on single-reviewer judgment with informal spot-checking. Numerical fields and effect-size calculations for the included quantitative records were also rechecked. The discrepancy identified in Asgari and Mal Amiri (2024) was resolved against the internally consistent source tables and reported t-test. On resolution, that study's comparison condition (self-directed comparison of learners' own output against a native-speaker text, with no explicit rule articulation or teacher-led grammar explanation) was judged not to meet this review's operational definition of a grammar-CR task, and the study was removed from the review during the citation-searching phase described below. This resolution does not alter the documented database coverage or the pooled estimates, but it is reported here in full for transparency.

The citation-searching stream carries its own limitations, documented fully in Appendix B. Twenty-six of the 28 seed reports were successfully forward-searched through Google Scholar's citation graph. Two seeds (Behrouzi & Kazemirad, 2012; Pan, 2008) were never indexed as cited by any retrievable work, and a third (Tashmuradova et al., 2023) became temporarily inaccessible on Google Scholar partway through; the search was paused at that point and later resumed and completed (Appendix B.3). Eight further records identified through the other-methods stream could not be retrieved despite reasonable effort and are therefore not reflected in the synthesis (full list and reasons in Appendix B.4). None of these eight records were estimated, approximated, or otherwise assumed for the synthesis; they are reported instead as an explicit, unresolved boundary on the completeness of the evidence base. The evidence base itself also imposes constraints. The meta-analysis combined different grammatical targets, instructional durations, comparison conditions, and outcome formats. Several studies relied on intact classes, yet intraclass correlations were unavailable for a consistent design-effect correction. Some standard deviations had to be reconstructed from reported standard errors or t-statistics, and gain scores or adjusted post-test estimates were used when raw endpoint dispersion was unavailable; two studies that reported gain scores were therefore excluded from the primary post-test-only synthesis and reported instead as a sensitivity analysis (Section 5.5). Many outcomes were immediate and closely aligned with the taught material, while delayed and spontaneous measures were uncommon. With 20 studies in the primary synthesis (22 in the sensitivity synthesis), regional comparisons, moderator analyses, funnel asymmetry, and publication-bias diagnostics remain exploratory rather than confirmatory.

9. Conclusion

This systematic review and meta-analysis provides focused quantitative evidence on grammar consciousness-raising in L2 English. Across the 20 controlled studies pooled in the primary post-test-only synthesis (1,317 learners), CR produced a large average advantage over non-CR comparison instruction ($g = 0.99$, 95% CI [0.78, 1.20]), a conclusion that held when two further gain-score studies were added as a sensitivity check ($k = 22$, $g = 1.03$). The prediction interval remained positive, and every leave-one-study-out estimate favored CR. The central finding is therefore not dependent on a single unusually positive study: asking learners to analyze linguistic evidence, compare forms, formulate generalizations, and test hypotheses can improve their performance on grammar-related outcomes.

The practical meaning of this result is more specific than the numerical magnitude alone suggests. CR appears most defensible as an analytic phase within a broader instructional sequence. Teachers can first use carefully selected examples to create a genuine grammatical problem. Learners then justify a rule or contrast, check that account against new data, and apply the feature in structured and communicative production. Used in this way, CR offers an alternative to both unsupported discovery and passive rule transmission. It gives learners intellectual responsibility for form while preserving a later role for practice, feedback, and meaning-focused use.

The review also establishes an important boundary on interpretation. Most included studies assessed immediate, controlled, or form-focused performance. Such outcomes provide evidence that learners can recognize, explain, or deliberately use the target feature; they do not by themselves demonstrate automatic access, implicit knowledge, durable retention, or spontaneous

communicative transfer. The results therefore support CR as a means of strengthening explicit grammatical understanding and controlled accuracy, not as a stand-alone guarantee of acquisition across all contexts and skills.

For research, the next priority is not simply to conduct more small comparisons using closely matched post-tests. Stronger primary trials should preregister their protocols, report assignment and attrition fully, adjust for classroom clustering, and include delayed tests, untrained items, timed measures, and spontaneous oral or written production. Future evidence syntheses should use independent duplicate screening and transparent extraction. Direct comparisons should also isolate which components of CR (data selection, collaborative reasoning, rule articulation, hypothesis testing, or subsequent practice) produce the observed benefit. Until that evidence accumulates, the most useful conclusion is appropriately bounded. Grammar CR is a promising and generally effective tool for developing conscious control of L2 English form, particularly when it is integrated with opportunities to proceduralize and communicate that knowledge.

Appendix A. Criterion-Level MMAT Appraisal

Randomized-trial criteria: 2.1 = Was randomization appropriately performed? 2.2 = Were the groups comparable at baseline? 2.3 = Are outcome data complete? 2.4 = Were outcome assessors blinded to the intervention? 2.5 = Did participants adhere to the assigned intervention? Non-randomized-study criteria: 3.1 = Are participants representative of the target population? 3.2 = Are measurements appropriate for the outcome and intervention? 3.3 = Are outcome data complete? 3.4 = Are confounders accounted for in the design and analysis? 3.5 = Was the intervention administered as intended?

Table 3. Criterion-level MMAT appraisal by study design (no overall numerical score).

Study	MMAT category	Criterion 1	Criterion 2	Criterion 3	Criterion 4	Criterion 5	Allocation evidence / principal concern
Amirian & Sadeghi (2012)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Two-group school study; random allocation to conditions was not reported.
Heidari & Mansourzadeh (2014)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Participants were divided after proficiency screening; random allocation was not reported.
Behrouzi & Kazemirad (2012)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	The report states that the 60 learners were randomly distributed into two groups (30/30), but the allocation-generation and concealment procedures were not described.
Idek et al. (2014)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 N	3.5 Y	Assignment to conditions was reported without a random-allocation method; baseline imbalance and control-group regression were present.
Abdalla (2014)	Non-randomized study	3.1 CT	3.2 CT	3.3 Y	3.4 CT	3.5 Y	The sample was randomly drawn, but random allocation to intervention and control was not documented.
Fatemipour & Hemmati (2015)	Randomized trial	2.1 CT	2.2 Y	2.3 Y	2.4 CT	2.5 Y	Random assignment to two groups was reported; sequence generation, concealment, and assessor blinding were not described.

Study	MMAT category	Criterion 1	Criterion 2	Criterion 3	Criterion 4	Criterion 5	Allocation evidence / principal concern
Tilahun et al. (2022)	Non-randomized study	3.1 N	3.2 Y	3.3 Y	3.4 Y	3.5 Y	A non-equivalent pretest–posttest design used intact sections selected and labeled by lottery.
Dang & Nguyen (2013)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Group assignment was not documented as random.
Tashmuradova et al. (2023)	Non-randomized study	3.1 N	3.2 Y	3.3 Y	3.4 Y	3.5 Y	Three intact classes formed a convenience sample; adjusted standard errors were used to reconstruct SD-equivalents.
Nosratinia & Roustayi (2014)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Learners were divided after homogenization; random allocation was not reported.
Qi (1994)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Very small intervention arm; two controls combined; allocation was not documented as random.
Yarahmadzahi et al. (2015)	Non-randomized study	3.1 N	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Four intact classes formed two intervention and two control groups.
Khezrlou (2024)	Randomized trial	2.1 CT	2.2 Y	2.3 Y	2.4 CT	2.5 Y	Random assignment to one control and two experimental groups was reported; the two CR arms were combined for independence.
Ma & Cho (2019)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Three intact classes were compared; the final immediate target test was selected by the outcome hierarchy.
Hassanzadeh & Shahbazi (2021)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Intact groups compared; random allocation to conditions was not reported.
Öncü (1998)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Independent-samples comparison of equal-sized experimental and control groups; random-allocation method not documented in the dissertation.

Study	MMAT category	Criterion 1	Criterion 2	Criterion 3	Criterion 4	Criterion 5	Allocation evidence / principal concern
Al-Lawatiyah (2012)	Non-randomized study	3.1 CT	3.2 Y	3.3 N	3.4 CT	3.5 Y	Enrolled sample (25 experimental/22 control) fell to an analyzed sample of 22/18 after attrition; random allocation not documented.
Brender (2002)	Non-randomized study	3.1 N	3.2 CT	3.3 Y	3.4 CT	3.5 Y	Unequal group sizes (56 experimental/35 control) are consistent with intact-class assignment rather than randomization.
Pakbaz & Rezai (2015)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Two intact groups compared; random-allocation method not documented.
Rahayu (2018)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Quasi-experimental design with a CR-versus-traditional-teaching control group; random-allocation method not documented.
Iskandar & Heriyawati (2015)	Non-randomized study	3.1 CT	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Quasi-experimental design; independent-samples t-test independently recomputed from reported means/SDs and matched the reported value exactly, supporting data integrity; random-allocation method not documented.
Al-Mubireek (2025)	Non-randomized study	3.1 Y	3.2 Y	3.3 Y	3.4 CT	3.5 Y	Quasi-experimental design (intact classes, same teacher); pretest equivalence confirmed (t-test, ns)

Note. Y = Yes (the criterion was met); N = No (the criterion was not met); CT = Can't tell (the report contained insufficient information for a definitive judgment). Criteria 2.1–2.5 apply only to randomized controlled trials, whereas Criteria 3.1–3.5 apply only to quantitative non-randomized studies. Random sampling was not treated as equivalent to random allocation. Intact- or cluster-class assignment was recorded as an allocation feature rather than as a separate MMAT design category. No overall MMAT score was calculated.

Appendix A.1. Risk-of-Bias Assessment (RoB 2 / ROBINS-I)

The MMAT appraisal above (Table 3) screens general methodological quality across mixed study designs but does not yield a structured risk-of-bias judgment for effect estimates. To address this directly, each of the 22 pooled studies was additionally appraised using the tool matched to its design: RoB 2 (Sterne et al., 2019) for the two studies with reported random assignment (Fatemipour & Hemmati, 2015; Khezlou, 2024), and ROBINS-I (Sterne et al., 2016) for the 20 non-randomized studies. Domain judgments were derived from the same source material used for the MMAT appraisal and for data extraction (allocation method, baseline-equivalence testing, attrition, and outcome measurement, as documented in Table 2 and Table 3). Where a domain's signaling questions could not be answered from the published report, the domain was recorded as "no information" (NI) rather than assigned a value not supported by the source text. Overall judgments follow each tool's standard algorithm, in which the least favorable domain determines the overall rating.

Table 4a. RoB 2 risk-of-bias judgments for the two randomized studies.

Study	D1 Randomization	D2 Deviations	D3 Missing data	D4 Outcome measurement	D5 Selective reporting	Overall
Fatemipour & Hemmati (2015)	Some concerns	Some concerns	Low	Some concerns	Some concerns	Some concerns
Khezrlou (2024)	Some concerns	Some concerns	Low	Some concerns	Some concerns	Some concerns

Table 4b. ROBINS-I risk-of-bias judgments for the 20 non-randomized studies.

Study	D1 Confounding	D2 Selection	D3 Classification	D4 Deviations	D5 Missing data	D6 Outcome measurement	D7 Selective reporting	Overall
Amirian & Sadeghi (2012)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Heidari & Mansourzadeh (2014)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Behrouzi & Kazemirad (2012)	Moderate	Low	Low	NI	Low	Moderate	NI	Moderate
Idek et al. (2014)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Abdalla (2014)	Serious	Low	Moderate	NI	Low	Moderate	NI	Serious
Tilahun et al. (2022)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Dang & Nguyen (2013)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Tashmuradova et al. (2023)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Nosratinia & Roustayi (2014)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Qi (1994)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Yarahmadzahi et al. (2015)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Ma & Cho (2019)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Hassanzadeh & Shahbazi (2021)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Öncü (1998)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Al-Lawatiyah (2012)	Serious	Moderate	Low	NI	Serious	Moderate	NI	Serious
Brender (2002)	Serious	Low	Moderate	NI	Low	Moderate	NI	Serious
Pakbaz & Rezai (2015)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Rahayu (2018)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Iskandar & Heriyawati (2015)	Serious	Low	Low	NI	Low	Moderate	NI	Serious
Al-Mubireek (2025)	Moderate	Low	Low	NI	Low	Moderate	NI	Moderate

Note. NI = no information. Domain 1 (confounding) was judged serious for 18 of the 20 non-randomized studies (no documented random allocation/concealment; no adjustment for pre-test scores) and moderate for two: Al-Mubireek (2025; pre-test equivalence confirmed on both outcomes, $p = .77/.96$, single instructor) and Behrouzi and Kazemirad (2012; random distribution reported, no significant pre-test difference, though allocation/concealment undescribed). Iskandar and Heriyawati (2015) was judged serious despite a randomization claim: it reported only a Levene's test ($p = .239$, which does not test mean differences), and pre-test means differed numerically (62.77 vs. 66.43) with no adjustment. Al-Lawatiyah (2012) was additionally serious on Domain 5 (attrition, 25/22 enrolled to 22/18 analyzed). Domain 6 was moderate throughout (objective measures, no reported blinding); Domain 7 was NI throughout (no registration). Both RCTs were rated some concerns overall.

Appendix B. Citation-Chasing Audit

This appendix expands on the citation-chasing summary in Section 3.1. Two separate counts of 28 occur in this review and should not be conflated. The first is the 28 seed reports (B.1), which comprised every eligible or candidate report already identified through the database and manual/supplementary search stages at the time citation-chasing began. The second is the 28 records shown in Figure 1 as identified from other methods, which combine the 10 of those seeds that came directly from the manual/supplementary search (rather than the formal database export) with the 18 further records that forward-searching all 28 seeds' citation graphs subsequently surfaced (17 new candidates and 1 duplicate).

B.1. The 28 seed reports. Eighteen seeds originated from the formal database search. These were Amirian and Sadeghi (2012); Heidari and Mansourzadeh (2014); Behrouzi and Kazemirad (2012); Nosratinia and Roustayi (2014); Idek, Fong, Sidhu, and Teoh (2014); Abdalla (2014); Fatemipour and Hemmati (2015); Tilahun, Simegn, and Emiru (2022); Dang and Nguyen (2013); Tashmuradova, Prasad, Abdel-Al Ibrahim, and Sarabani (2023); Qi (1994); Yarahmadzahi, Esfandiari Ghalee, and Kalali Sani (2015); Ma and Cho (2019); Khezrlou (2024); Öncü (1998); Al-Lawatiyah (2012); Brender (2002); and Asgari and Mal Amiri (2024),

the last later excluded for construct mismatch (Section 3.1). The remaining 10 seeds originated from the manual/supplementary search that predated formal citation-chasing. These were Pourcheragh (2019); Özbay (2016); Azizifar, Babaei, Jamalinesari, and Gowhary (2015); Hassanzadeh and Shahbazi (2021); Xavier and Gesser (2022); Hosseinik and Sangani (2014); Erfani and Torkamani (2015); Arshad et al. (2015); Amirian and Abbasi (2014); and Pan (2008).

B.2. Forward citation-searching procedure. On 9 August 2026, each of the 28 seed reports was located on Google Scholar, using a direct title search or, where a title search returned an ambiguous or unindexed result, a DOI-based lookup to confirm the correct record. Where a seed's Google Scholar entry showed a "Cited by" link, every citing work listed, across all result pages, was screened by title and, where available, abstract, against this review's eligibility criteria (Table 1). Those criteria were a grammar consciousness-raising intervention, a non-CR comparator, an L2-English morphosyntactic outcome, and a controlled or quasi-controlled design. Titles that plausibly matched were flagged for full-text retrieval and assessment. Titles that clearly did not match (wrong construct, wrong outcome domain, non-empirical work, or an already-processed candidate) were excluded at this screening step without further retrieval, consistent with the title/abstract screening logic applied to the database stream. A supplementary Web of Science Core Collection check was also attempted for four seeds after Google Scholar access was briefly disrupted (B.3); it returned negligible coverage of this review's literature and surfaced no new candidates, so it was not pursued further.

B.3. Seeds that could not be forward-searched. Twenty-six of the 28 seeds were successfully searched. Two seeds could not be searched via Google Scholar at all. Behrouzi and Kazemirad (2012) has a DOI that resolves to a genuine, verifiable article but returns no indexed citation record on Google Scholar, and Pan (2008), a Chinese-language dissertation, returned no matching record under any title, author, or institution search on Google Scholar. A third seed, Tashmuradova et al. (2023), became temporarily inaccessible on Google Scholar partway through the check. The search for this seed was paused at that point and resumed and completed later in the same session (5 citing works screened, 0 new candidates).

B.4. Records not retrieved ($n = 8$). Of the 19 other-methods records assessed for eligibility (12 retained, 7 excluded; B.5), 8 further records could not be retrieved despite reasonable effort and are not reflected in the synthesis. These were Amirian and Abbasi (2014), paywalled with no institutional access; Abedi and Ashari Tabar (2018), full text not locatable; Kamarulzaman (2018), ProQuest dissertation with no accessible route; Pei (n.d.), conference paper not locatable; Panarook (2022), abstract retrieved but PDF download failed repeatedly; Farooq, Zeeshan, and Palwasha (2025), not retrievable in either publication venue; Nugroho (2018), PDF link broken; and Hashemi Saeid et al. (2021), bare citation record only. None of these eight records were estimated, approximated, or otherwise assumed for the synthesis.

B.5. Records excluded after full-text assessment ($n = 7$). Four were excluded for lacking any control or comparison group. These were Gobena (2022), a descriptive survey with no experimental design; Suhaimi and Musdizal (2022), a classroom action-research study with no comparison group; Handoyo, Hariyanto, Widiyowati, and Mustain (2023), a one-group pre-experimental design by the authors' own description; and Abay (2021), a single-group interrupted time-series design. Two were excluded for construct mismatch. Arshad et al. (2015) contrasted use versus non-use of the learners' first language as the medium of instruction, rather than an isolated CR-versus-no-CR contrast, and Behbahani and Khademi (2022) bundled CR with input flooding and visual input enhancement in a design where the comparison group received no instruction on the target structure at all, so that CR could not be isolated as the differentiating variable. One further seed report, Pan (2008), was excluded on accessibility grounds distinct from the eight records in B.4. Its ProQuest record explicitly confirms zero pages of full text available through any channel (citation and abstract only), which this review treats as a definitive rather than a pending inaccessibility, unlike the eight records in B.4, where a full text plausibly exists but was not secured this session.

B.6. Supplementary post-review search (10 August 2026). In response to peer-review feedback that the original search strategy (Sections III.A and B.1-B.2) may have under-retrieved eligible records, a targeted supplementary search was conducted using broader consciousness-raising terms than the original strategy (including "grammatical consciousness raising," "grammar awareness raising," and "GCRT*"), together with direct verification of two specific records raised in review. Two records were identified and read in full. Al-Mubireek (2025), a quasi-experimental comparison of Grammar Consciousness-Raising Tasks against traditional grammar instruction for first-year Saudi EFL university students, met this review's eligibility criteria and was added to the strict grammar synthesis ($k: 21 \rightarrow 22$; $N = 1,473$; Table 2). Moradkhan and Sohrabian (2009), despite its title, was assessed and excluded for construct mismatch: the group it labels "consciousness-raising" in fact received teacher-fronted explicit grammar explanation before practice, which this review's operational definition of CR (Section 2.1) excludes, while its comparison group received purely communicative activities with no explicit form focus. This supplementary check targeted the two records raised in review rather than constituting a full re-run of the search strategy across all databases; it is reported here for the same reason the original search is reported in full, consistent with this review's no-fabrication and full-transparency principle.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Protocol and Data Availability: The search log, extraction dataset, and analysis code accompany this manuscript and are available from the author upon request.

ORCID iD: 0000-0003-2071-1734

Declaration of the Use of AI: The author declares that this work was prepared entirely without the use of generative AI technologies or automated text-producing tools.

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

- [1] Abdalla, A. H. (2014). Investigating the effect of consciousness-raising techniques on learning the English tense system by EFL students. *Journal of Language Teaching and Research*, 5(4), 810-818. <https://doi.org/10.4304/jltr.5.4.810-818>
- [2] Al-Lawatiyah, S. (2012). Consciousness-raising tasks in EFL: Design and impact (Doctoral dissertation, Sultan Qaboos University). ProQuest Dissertations & Theses Global (305576795).
- [3] Al-Mubireek, S. (2025). Impact of grammar consciousness-raising tasks on the acquisition of English tenses at university level. *Evolutionary Studies in Imaginative Culture*, 9(2), 263-275. <https://doi.org/10.70082/esiculture.vi.3077>
- [4] Amirian, S. M. R., & Abbasi, S. (2014). The effect of grammatical consciousness-raising tasks on Iranian EFL learners' knowledge of grammar. *Procedia - Social and Behavioral Sciences*, 98, 228-235.
- [5] Amirian, S. M. R., & Sadeghi, F. (2012). The effect of grammar consciousness-raising tasks on EFL learners' performance. *International Journal of Linguistics*, 4(3). <https://doi.org/10.5296/ijl.v4i3.2392>
- [6] Asgari, L., & Mal Amiri, B. (2024). The impact of autonomous tasks on grammar achievement among EFL learners. *International Journal of Language Education and Applied Linguistics*, 14(1). <https://doi.org/10.15282/ijleal.v14i1.10081>
- [7] Azizifar, A., Babaei, M., Jamalinesari, A., & Gowhary, H. (2015). The effect of grammatical consciousness raising task on Iranian EFL learners' reading comprehension. *Procedia - Social and Behavioral Sciences*, 192, 252-259. <https://doi.org/10.1016/j.sbspro.2015.06.036>
- [8] Behrouzi, P., & Kazemirad, F. (2012). The effect of consciousness-raising tasks on Iranian elementary EFL learners' syntax acquisition. *International Journal of Linguistics*, 4(2). <https://doi.org/10.5296/ijl.v4i2.1852>
- [9] Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. <https://doi.org/10.1002/9780470743386>
- [10] Brender, A. (2002). The effectiveness of teaching articles to (-ART) ESL students in writing classes using consciousness-raising methods (Doctoral dissertation, Temple University). ProQuest Dissertations & Theses Global.
- [11] Dang, T. T., & Nguyen, H. T. (2013). Direct versus indirect consciousness-raising instruction in English grammar teaching. *English Language Teaching*, 6(1), 112-121. <https://doi.org/10.5539/elt.v6n1p112>
- [12] Doughty, C., & Williams, J. (Eds.). (1998). *Focus on form in classroom second language acquisition*. Cambridge University Press.
- [13] Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629-634. <https://doi.org/10.1136/bmj.315.7109.629>
- [14] Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, 27(2), 305-352. <https://doi.org/10.1017/S027226310505014X>
- [15] Ellis, R. (2001). Introduction: Investigating form-focused instruction. *Language Learning*, 51(S1), 1-46. <https://doi.org/10.1111/j.1467-1770.2001.tb00013.x>
- [16] Ellis, R. (2002). Grammar teaching—practice or consciousness-raising? In J. C. Richards & W. A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice* (pp. 167-174). Cambridge University Press. <https://doi.org/10.1017/CBO9780511667190.023>
- [17] Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, 27(2), 141-172. <https://doi.org/10.1017/S0272263105050096>
- [18] Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly*, 40(1), 83-107. <https://doi.org/10.2307/40264512>
- [19] Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, 28(2), 339-368. <https://doi.org/10.1017/S0272263106060141>
- [20] Erfani, S. M., & Torkamani, F. (2015). The impact of consciousness raising vs. structure-based production tasks on reading comprehension of Iranian EFL learners. *Theory and Practice in Language Studies*, 5(5). <https://www.academypublication.com/issues2/tpls/vol05/05/24.pdf>
- [21] Fatemipour, H., & Hemmati, S. (2015). Impact of consciousness-raising activities on young English language learners' grammar performance. *English Language Teaching*, 8(9), 1-10. <https://doi.org/10.5539/elt.v8n9p1>
- [22] Fazlali, B., & Shahini, A. (2019). The effect of input enhancement and consciousness-raising techniques on the acquisition of lexical and grammatical collocation of Iranian EFL learners. *TESL-EJ*, 23(2).
- [23] Fotos, S. S. (1993). Consciousness raising and noticing through focus on form: Grammar task performance versus formal instruction. *Applied Linguistics*, 14(4), 385-407. <https://doi.org/10.1093/applin/14.4.385>

- [24] Fotos, S., & Ellis, R. (1991). Communicating about grammar: A task-based approach. *TESOL Quarterly*, 25(4), 605-628. <https://doi.org/10.2307/3587079>
- [25] Hassanzadeh, M., & Shahbazi, F. (2021). Explicit instruction of English articles: An appraisal of consciousness-raising instruction and processing instruction frameworks. *Journal of AsiaTEFL*, 18(4).
- [26] Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107-128. <https://doi.org/10.3102/10769986006002107>
- [27] Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60-87. <https://doi.org/10.3102/0162373707299706>
- [28] Heidari, F., & Mansourzadeh, N. (2014). Grammar consciousness-raising activities and Iranian EFL learners' grammar performance. *Journal of Education and Learning*, 8(3). <https://doi.org/10.11591/edulearn.v8i3.18496>
- [29] Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327, 557-560. <https://doi.org/10.1136/bmj.327.7414.557>
- [30] Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT), version 2018: User guide. McGill University.
- [31] Hosseini, S. Y., & Sangani, H. R. (2014). The influence of explicit cross-linguistic consciousness-raising on the EL writing of the Iranian English language learners. *Advances in Language and Literary Studies*, 5(3), 147-152. <https://doi.org/10.7575/aialc.all.v5n.3p.147>
- [32] Idek, S., Fong, L. L., Sidhu, G. K., & Teoh, S. H. (2014). The use of consciousness-raising tasks in promoting the correct use of the verb "be" among students in vocational colleges. *Theory and Practice in Language Studies*, 4(9), 1792-1803. <https://doi.org/10.4304/tpls.4.9.1792-1803>
- [33] Int'Hout, J., Ioannidis, J. P. A., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, 14, 25. <https://doi.org/10.1186/1471-2288-14-25>
- [34] Iskandar, D., & Heriyawati, D. F. (2015). Grammar consciousness-raising activities and their impact on students' grammatical competence. *JEELS (Journal of English Education and Linguistics Studies)*, 2(2).
- [35] Khezrlou, S. (2024). Effects of task repetition with consciousness-raising in wiki-mediated collaborative writing on the development of explicit and implicit knowledge. *Computer Assisted Language Learning*, 37(1-2), 243-278. <https://doi.org/10.1080/09588221.2022.2033789>
- [36] Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware versus unaware learners. *Studies in Second Language Acquisition*, 22(4), 557-584. <https://doi.org/10.1017/S0272263100004046>
- [37] Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, 60(2), 309-365. <https://doi.org/10.1111/j.1467-9922.2010.00561.x>
- [38] Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition*, 27(3), 361-386. <https://doi.org/10.1017/S0272263105050163>
- [39] Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. B. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39-52). John Benjamins. <https://doi.org/10.1075/sibil.2.07lon>
- [40] Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition*, 26(3), 399-432. <https://doi.org/10.1017/S0272263104263021>
- [41] Ma, J. H., & Cho, Y. A. (2019). The effects of consciousness-raising tasks on EFL learners' grammar acquisition and flow. *Journal of Research in Curriculum & Instruction*, 23(1), 16-25. <https://doi.org/10.24231/rici.2019.23.1.16>
- [42] Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, 27(3), 405-430. <https://doi.org/10.1093/applin/ami051>
- [43] Moradkhan, D., & Sohrabian, R. (2009). Grammatical consciousness-raising tasks and the improvement of Iranian students' knowledge of grammar. *Journal of Teaching English as a Foreign Language and Literature*, 1(4), 57-71.
- [44] Namaghi, S. A. O., & Charmchi, M. (2016). Grammar consciousness-raising tasks and intermediate learners' achievement and perceptions. *Sino-US English Teaching*, 13(1), 1-7.
- [45] Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417-528. <https://doi.org/10.1111/0023-8333.00136>
- [46] Nosratinia, M., & Roustayi, S. (2014). The effect of grammar consciousness-raising tasks on EFL learners' reading comprehension and writing ability. *IOSR Journal of Humanities and Social Science*, 19(3), 203-216. <https://doi.org/10.9790/0837-1932203216>
- [47] Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- [48] Pakbaz, R., & Rezai, M. J. (2015). The effect of task repetition and consciousness-raising on Iranian L2 learners' writing performance. *International Journal of Language and Applied Linguistics*, 1(2), 58-70.

- [49] Panarook, C. (2022). The effectiveness of consciousness-raising tasks in teaching and learning phrases and clauses of L2 learners at tertiary level. *Journal of Studies in the Field of Humanities*, 29(2), 223-252.
- [50] Pourcheragh, M. (2019). The effect of consciousness-raising instruction on the pragmatic development of agreeing and disagreeing of Iranian EFL upper intermediate learners in speaking performance. *Revista San Gregorio*, 36. <https://doi.org/10.36097/rsan.v0i36.1016>
- [51] Qi, D. S. (1994). Consciousness raising in ESL writing: Its effects on the syntactic skills and writing proficiency of university students (Doctoral dissertation). ProQuest Dissertations & Theses Global (9429844).
- [52] Rahayu, L. (2018). The effectiveness of using GC-raising activities on the student's grammar score of the first graders of MA Ar-Rosyaad Balong Ringinrejo (Undergraduate thesis, IAIN Kediri). <https://etheses.iainkediri.ac.id/15944/>
- [53] Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, 342, d549. <https://doi.org/10.1136/bmj.d549>
- [54] Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning*, 45(2), 283-331. <https://doi.org/10.1111/j.1467-1770.1995.tb00441.x>
- [55] Rutherford, W. E., & Sharwood Smith, M. (1985). Consciousness-raising and Universal Grammar. *Applied Linguistics*, 6(3), 274-282. <https://doi.org/10.1093/applin/6.3.274>
- [56] Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129-158. <https://doi.org/10.1093/applin/11.2.129>
- [57] Sharwood Smith, M. (1981). Consciousness-raising and the second language learner. *Applied Linguistics*, 2(2), 159-168. <https://doi.org/10.1093/applin/II.2.159>
- [58] Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, 60(2), 263-308. <https://doi.org/10.1111/j.1467-9922.2010.00562.x>
- [59] Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., ... Higgins, J. P. T. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, 355, i4919. <https://doi.org/10.1136/bmj.i4919>
- [60] Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., ... Higgins, J. P. T. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*, 366, l4898. <https://doi.org/10.1136/bmj.l4898>
- [61] Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64-81). Cambridge University Press.
- [62] Tashmuradova, B., Prasad, K. D. V., Abdel-Al Ibrahim, K. A., & Sarabani, L. (2023). The comparative effects of consciousness-raising tasks and dynamic assessment on morphological awareness: The case of pre-intermediate EFL learners. *Language Testing in Asia*, 13, Article 27. <https://doi.org/10.1186/s40468-023-00231-7>
- [63] Tilahun, S., Simegn, B., & Emiru, Z. (2022). Using grammar consciousness-raising tasks to enhance students' narrative tenses competence. *Cogent Education*, 9(1), Article 2107471. <https://doi.org/10.1080/2331186X.2022.2107471>
- [64] Xavier, R. P., & Gesser, A. R. (2022). The role of a consciousness-raising task in a focused-task sequence. *Revista da Anpoll*, 53(1), 78-98. <https://doi.org/10.18309/ranpoll.v53i1.1610>
- [65] Yarahmadzahi, N., Esfandiari Ghalaee, A., & Kalali Sani, S. F. (2015). The effect of teaching grammar through consciousness raising tasks on high school English learners' grammatical proficiency. *Mediterranean Journal of Social Sciences*, 6(3 S1), 401-408. <https://doi.org/10.5901/mjss.2015.v6n3s1p401>
- [66] Zhang, R. (2015). Measuring university-level L2 learners' implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition*, 37(3), 457-486. <https://doi.org/10.1017/S0272263114000370>
- [67] Öncü. (1998). A study on comparison of grammar consciousness-raising tasks and traditional teacher-fronted grammar instruction on teaching modals in English (Doctoral dissertation, Anadolu University). ProQuest Dissertations & Theses Global (28640500).
- [68] Özbay, A. S. (2016). The potential of corpus-based collocation instruction on the awareness levels of the Turkish EFL students in terms of reading performances. In A. Moreno Ortiz & C. Pérez-Hernández (Eds.), *CILC2016: 8th International Conference on Corpus Linguistics* (Vol. 1, pp. 331-340). EasyChair. <https://doi.org/10.29007/k2lr>