從考生到出題者：透過學生生成的文法評量監測翻轉英語作為外語課堂中的認知與反思發展

楊晴絨

南臺科技大學雙語教學推動中心

cjyang@stust.edu.tw

摘要

本研究採設計導向方法，探討學生生成文法評量如何促進翻轉英語作為外語（EFL）課堂中的認知複雜性、評估素養與後設認知發展。研究歷時兩學期，共有 116 人次來自南台灣某科技大學應用英語系的學生參與學生生成評估（SGA）循環，協作設計、執行並反思與課程目標相符的文法測驗。研究以布魯姆修訂分類法、Swain 的產出假設及 Fulcher 的評估素養框架為理論依據，分析 685 份學習者撰寫的題目、反思報告、同儕評估及前後測成績。結果顯示，高階思維任務比例顯著提升：應用至創造層級的題目由 66.4%增至 89.6%，錯誤分析與句子重構題型亦有明顯增長。學習者在題目有效性、構念對準及干擾項設計方面展現更深理解，反思中常採用「教師視角」進行分析。各組文法能力顯著提升，學習者普遍認為進步來自反覆設計歷程與同儕互評機制。質性資料揭示評估創作的情感與人際面向，展現學習者在創造力與理解度、自主性與責任感間的平衡。SGA 循環將文法學習從規則記憶轉化為參與式探究，使學習者成為教學的貢獻者，而非被動的受試者。本研究確認 SGA 為一具可行性與影響力的教學策略，能深化文法理解並促進反思主體性，提供一個可複製的模型，將評估素養有效融入 EFL 教學，兼顧學習者聲音與教學嚴謹性。

關鍵詞：學生生成評估、文法教學、認知複雜性、評估素養、翻轉英語作為外語（EFL）課堂

From Test-Takers to Test-Makers: Monitoring Cognitive and Reflective Development Through Student-Generated Grammar Assessment in a Flipped EFL Classroom

Ching-Jung Yang

Center for Bilingual Education, Southern Taiwan University of Science and Technology

Abstract

This design-based study explores how student-generated grammar assessments can cultivate cognitive complexity, assessment literacy, and meta-cognitive growth within a flipped English as a Foreign Language (EFL) classroom. Over two academic semesters, 116 Applied English majors at a technical university in southern Taiwan engaged in a Student-Generated Assessment (SGA) cycle, collaboratively designing, implementing, and reflecting on grammar quizzes aligned with instructional goals. Guided by Bloom's Revised Taxonomy, Swain's Output Hypothesis, and Fulcher's assessment literacy framework, the study examined 685 learner-authored test items,

Received: Aug. 11, 2025; first revised: Sep. 2, 2025; accepted: Sep. 2025.

Corresponding author: C.-J. Yang, Center for Bilingual Education, Southern Taiwan University of Science and Technology, Tainan 710301, Taiwan

reflective narratives, peer evaluations, and pre-/post-test data. Results revealed a marked shift toward higher-order thinking: tasks at the Apply through Create levels increased from 66.4% to 89.6%, with substantial growth in error analysis and sentence reconstruction formats. Learners demonstrated heightened sensitivity to item validity, construct alignment, and the plausibility of distractors, frequently adopting a "teacher's lens" in their reflections. Across cohorts, grammar proficiency improved significantly, with students attributing their progress to the iterative nature of the design process and the critical engagement fostered by peer review. Qualitative insights underscored the emotional and interpersonal dimensions of assessment authorship—highlighting the delicate balance between creativity and clarity, agency and accountability. By transforming grammar from a rule-based exercise into a participatory inquiry, the SGA cycle repositioned learners as active contributors to instruction rather than passive recipients of evaluation. This study affirms SGA as a feasible and impactful pedagogical strategy for deepening grammatical understanding and nurturing reflective agency. It offers a replicable model for embedding assessment literacy into EFL instruction—one that amplifies learner voice while upholding academic rigor.

Keywords: Student-generated Assessment, Grammar Instruction, Cognitive Complexity, Assessment Literacy, Flipped EFL Classroom

I. Introduction

In Taiwanese tertiary education, English majors must navigate the dual demands of standardized proficiency benchmarks and the communicative competencies required for careers in teaching, translation, and global business (Hung & Huang, 2019; Lin, 2016). Grammar instruction, however, often remains anchored in decontextualized drills and teacher-led explanations—approaches that have been criticized for stifling creativity, limiting learner agency, and constraining higher-order cognition (Li & Wilson, 2025; Pawlak, 2024).

Flipped-classroom models offer partial relief by relocating lower-level input outside class, thereby freeing contact time for active use (Zainuddin & Halili, 2016). Yet even flipped lessons can default to teacher-controlled quizzes that perpetuate passive assessment paradigms.

Student-generated assessment (SGA) represents a more radical pedagogical shift. By positioning students as co-designers of test items, reviewers of peer work, and interpreters of statistical evidence, SGA transforms assessment into a participatory and reflective process (Black & Wiliam, 2018). Prior studies have documented achievement gains when learners construct multiple-choice questions (Katz et al., 2024) or generate feedback on peer-authored items (Yu & Wu, 2017). These practices activate metacognitive and self-regulatory processes essential to long-term learning (Panadero et al., 2017). However, most empirical work treats quiz creation as a one-off homework task, leaving unanswered questions about how sustained cycles of collaborative design, peer delivery, and reflective reporting shape cognitive complexity and assessment literacy over time.

This study addresses these gaps by zooming in on the student-generated assessment strand embedded within a broader Tripartite Quiz-Based Flipped Classroom (TQB-FC) intervention—a three-phase instructional framework integrating self-learning, peer learning, and hands-on learning activities (detailed in Section III; currently under review for publication). While the companion paper ("Scaffolded Quiz Ecologies in Flipped Learning") demonstrated robust grammar-proficiency gains (d = 0.85~0.90) and metacognitive growth, it did not disentangle which components of the framework drove these improvements. Here, the analytic lens is narrowed to the assessment work students themselves performed: generating quiz items of varied formats, interpreting facility indices (i.e., item difficulty scores indicating the proportion of correct responses), drafting collaborative

reports, and evaluating peers' quizzes, with corresponding survey measures focused specifically on these SGA processes.

Guided by a theoretical framework that integrates sociocultural theory, Bloom's Revised Taxonomy, and assessment literacy constructs, this study explores the following objectives:

1. Trace changes in the cognitive level (Bloom-aligned) of student-generated grammar questions across two semesters.
2. Document evidence of assessment literacy and metacognitive development in collaborative reports, peer-evaluation comments, and interviews.
3. Gauge changes in learner satisfaction with peer-created quizzes.
4. Consider whether pre-/post-grammar tests can serve as indirect indicators of improved assessment literacy.

By examining these dimensions, the study contributes to a growing body of research that reimagines assessment not as a static endpoint, but as a dynamic and dialogic process—one that empowers learners to take ownership of their knowledge and its evaluation.

II. Literature Review

This study investigates SGA as a pedagogical strategy for enhancing grammar learning, assessment literacy, and cognitive engagement in an English as a Foreign Language (EFL) context. SGA reframes learners not merely as recipients of evaluation but as co-constructors of assessment, engaging them in the design, critique, and reflection of learning tasks. The literature review is organized into four interrelated strands: (1) empirical findings on learner-authored assessments, (2) cognitive-complexity frameworks, (3) collaborative assessment literacy, and (4) identified research gaps.

1. Student-Generated Assessment and Language Learning

SGA positions learners as active agents in their educational journey, encouraging them to design test items that reflect their understanding and challenge their peers. This approach aligns with constructivist and sociocultural theories, which emphasize learning through production, negotiation, and reflection.

Recent studies affirm the cognitive and motivational benefits of learner-authored assessments. Liu et al. (2025) examined an automated corrective feedback-based peer assessment model in foreign language pronunciation courses, finding that students who engaged in reflective peer review and correction demonstrated significantly higher achievement, intrinsic motivation, and self-regulated learning conceptions. Their findings suggest that when learners engage in designing or critiquing assessment tasks, they engage more deeply with linguistic input and output processes.

In a flipped EFL grammar context, Xia et al. (2024) conducted a scoping review on how generative AI transforms assessment practices. They found that student-generated grammar tasks—especially those supported by AI tools—enhanced learners' metacognitive awareness and fostered more responsible, self-directed learning. These environments encouraged students to reflect on grammatical structures, anticipate peer misunderstandings, and co-construct meaning through collaborative item design.

These findings echo earlier work by Panadero (2017), whose meta-analysis concluded that student-generated assessment fosters self-regulated learning by activating planning, monitoring, and evaluative processes. When learners design tasks, they must clarify their understanding, anticipate cognitive demands, and reflect on how knowledge is demonstrated—processes that mirror core dimensions of metacognition. Moreover, collaborative item design encourages dialogic learning, where students co-construct meaning and critique each other's reasoning. Extending this perspective, Brandmo et al. (2020) emphasize that formative assessment can serve as a

catalyst for self-regulated learning when it supports the full cycle of forethought, performance, and self-reflection. Their framework also highlights the role of co-regulation and socially shared regulation, suggesting that assessment tasks embedded in collaborative contexts—such as SGA—can deepen learners' strategic engagement and foster shared cognitive responsibility.

2. Cognitive-Complexity Frameworks

The quality of learner-generated items is often evaluated through cognitive-complexity frameworks, with Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) serving as the predominant model. Bloom's taxonomy provides a hierarchical lens for analyzing the depth of cognitive engagement, ranging from lower-order skills (e.g., remembering, understanding) to higher-order processes (e.g., applying, analyzing, evaluating, and creating).

Recent literature has expanded this framework to include integrative complexity and cognitive load theory. Hernandez Sibo et al. (2024) synthesized 33 studies on cognitive load in creative thinking, highlighting how task design influences learners' ability to balance novelty and usefulness—an insight relevant to grammar item construction. Meanwhile, Molina et al. (2023) emphasized the role of integrative complexity in decision-making and assessment, suggesting that learners who engage in item design exhibit higher levels of differentiation and integration in their cognitive processing.

In grammar-focused contexts, Gebregziabher et al. (2025) demonstrated that flipped learning environments scaffold higher-order grammar reasoning, particularly when students are tasked with creating and evaluating peer-generated items. These findings support the use of simplified Bloom-based rubrics for coding learner-generated content, as adapted in the present study to ensure inter-rater reliability.

3. Assessment Literacy in Collaborative Contexts

Assessment literacy refers to the knowledge, skills, and dispositions required to understand, interpret, and use assessment effectively. Fulcher (2012) conceptualizes it as encompassing technical knowledge (e.g., validity, reliability), critical principles (e.g., fairness, transparency), and socio-ethical awareness (e.g., inclusivity, learner empowerment).

Recent scholarships have emphasized the collaborative dimensions of assessment literacy. Meijer et al. (2020) introduced the concept of "collaborative learning assessment literacy," arguing that group-based assessment tasks—when paired with peer feedback and teacher scaffolding—enhance learners' understanding of assessment purposes and consequences. Mphahlele (2024) reviewed 38 studies in open-distance learning contexts, concluding that collaborative assessments foster engagement, motivation, and shared responsibility. However, as Pastore (2023) notes in her systematic review, assessment literacy in higher education remains under-conceptualized, with definitional ambiguity and limited integration into instructional practice. These gaps underscore the need for clearer frameworks and learner-centered models that promote both technical and reflective dimensions of assessment.

In the present study, assessment literacy is cultivated through multiple channels: quiz design, peer evaluation, reflective reporting, and survey-based feedback. These activities invite students to interrogate the grammar proficiency—measured via pre-/post-tests—might serve as indirect indicators of enhanced assessment insight.

4. Gaps and Rationale

Despite growing interest in student-generated assessment, several gaps persist in the literature. Addressing these gaps is timely for curricula seeking to reconcile exam-oriented demands with learner-centered, higher-order pedagogy. The present study contributes to this effort by implementing a semester-long SGA cycle that integrates

quiz design, collaborative reflection, and survey-based inquiry—positioning students not only as test-takers but also as thoughtful designers and evaluators of their own learning. By tracing changes in cognitive complexity, documenting assessment literacy development, and capturing learner perceptions, the study offers a comprehensive portrait of how student-generated assessment can transform grammar instruction into a participatory, reflective, and intellectually rigorous practice.

III. Methodology

This study employed a design-based research (DBR) approach to examine how student-generated assessments embedded within a TQB-FC framework foster grammar learning, cognitive complexity, and assessment literacy among Taiwanese EFL learners. DBR was selected for its iterative, context-sensitive nature, allowing pedagogical theory to be assessed and refined through authentic classroom practice (McKenney & Reeves, 2018).

1. Participants and Setting

A total of 116 first-year Applied English majors enrolled in the required courses *English Grammar and Rhetoric I and II* at a private technical university in southern Taiwan participated across two academic semesters. The fall cohort comprised 61 students, and the spring cohort 55, with 55 students completing both semesters. The overall attrition rate was 9.8% (n = 6), primarily due to academic program transfers.

Participants were predominantly aged 18–20 (82%), with the remainder consisting of older transfer or repeat students. The cohort included 57 domestic students and 4 international students from Malaysia, Vietnam, and overseas Chinese communities. All international participants held A2-level Chinese proficiency certification, enabling bilingual instructional support throughout the intervention.

Baseline English proficiency scores, drawn from the technical college entrance examination, averaged 78 for general English and 60 for English reading and writing. These diagnostic measures informed the implementation of differentiated scaffolding. To ensure balanced peer interaction and support, heterogeneous teams of 4~5 students were formed through stratified randomization based on initial proficiency levels.

Instruction took place in a well-equipped language learning lab featuring six desktop computers, two ceiling-mounted projectors, and movable tables and chairs. This flexible setup supported dynamic transitions between individual and collaborative work. Students accessed Google Docs and the FlipClass platform, the school's learning management system (LMS), via lab computers or personal devices. While digital autonomy was encouraged for most tasks, quiz sessions emphasized peer interaction and academic integrity by requiring device-free collaboration in assigned teams. The setting reflected the participatory and tech-integrated ethos of the instructional design.

2. Instructional Content and Grammar Focus

Each semester's instructional sequence followed the TQB-FC framework and was designed to foster progressive cognitive engagement through grammar instruction. Drawing on *Grammar in Context 2* by Sandra N. Elbaum (Elbaum, 2016), the Fall Semester emphasized foundational grammar forms such as present and past tenses, possessives, pronouns, modifiers, and quantity expressions. The Spring Semester introduced more complex structures, including modals, perfect and continuous aspects, gerunds and infinitives, adjective clauses, voice distinctions, and article usage. These content choices, aligned with curriculum standards and learner proficiency profiles, informed the cognitive demands of student-generated quiz items.

3. Instructional Intervention: Tripartite Quiz-Based Flipped Classroom (TQB-FC)

The instructional intervention was structured around the TQB-FC framework, which sequenced learning across three interdependent phases—Self-Learning (SL), Peer Learning (PL), and Hands-On Learning (HL). Each phase was designed to scaffold grammar acquisition, promote learner agency through quiz authorship, and accommodate diverse learner profiles via technology-mediated and paper-based modalities.

(1) Self-Learning (SL)

In the self-learning phase, students engaged with narrated micro-lectures developed by the instructor, supplemented by textbook-based grammar exercises, curated multimedia content, and structured posting tasks via the FlipClass platform. To support low-stakes self-assessment, answer keys to selected textbook exercises were posted on FlipClass, enabling students to check their own work independently. Weekly video materials (30~40 minutes) were designed to activate prior knowledge and promote grammar noticing in preparation for in-class peer interaction. In the fall semester, emoji reactions in the FlipClass Forum were encouraged to foster informal peer engagement; however, this practice was discontinued in the spring due to limited reflective depth and inconsistent interpretability.

(2) Peer Learning (PL)

In-class sessions featured printed paper-based quizzes completed individually and then collaboratively within designated teams. External resources were restricted to promote in-group reasoning. Quiz-giving teams collected annotated responses, verified answers, and prepared feedback materials. Instructor and teaching assistant (TA) support focused on scaffolding for lower-proficiency learners and clarifying contestable items. Oral discussion tasks were reconceptualized as written mistake analyses submitted via FlipClass, enhancing metalinguistic awareness and differentiated support.

(3) Hands-On Learning (HL)

To scaffold quiz design, the instructor introduced model quizzes featuring cognitively demanding “trap” items that integrated vocabulary and reading content. Student groups contributed printed quizzes or Kahoot-based games one to two times per semester. Drafts were reviewed via FlipClass, with flawed items sometimes retained to prompt peer-level analysis. Quiz-giving teams also completed their own quizzes to ensure reciprocal engagement and led review discussions in subsequent sessions. These were reinforced by FlipClass Forum summaries and immediate feedback in Kahoot sessions. Collaborative quizzes authoring via Google Docs supported real-time negotiation and learner autonomy.

4. Student-Generated Assessment (SGA) Cycle

Within the broader TQB-FC framework, the present study focused specifically on the SGA strand, implemented as a semester-long cycle to examine how learners engaged with grammar content through collaborative quiz design, reflective reporting, and peer evaluation. Although Kahoot was initially considered as a platform for quiz-based activities, it was excluded from the final analysis due to its limited functionality—specifically, its restriction to quiz-type questions without support for open-ended or multi-modal formats. These constraints were misaligned with the study’s emphasis on learner agency and deeper metalinguistic engagement.

(1) Collaborative Assessment Design

Students worked in heterogeneous teams to design grammar-in-context quizzes aligned with assigned topics, totaling 100 points and incorporating varied item types (e.g., multiple-choice, fill-in-the-blank, and error correction). Quizzes were co-authored via Google Docs and submitted two weeks prior to administration. Instructor feedback addressed topic alignment and surface-level issues, while pedagogically flawed items were

intentionally retained to prompt in-class discussions on item validity and cognitive demand.

(2) Dual-Phase Quiz Administration

Quizzes were administered in class using a paper-based format. Students first completed the quiz individually (trial phase), followed by a collaborative group phase in which teams negotiated responses and submitted a consensus version. This structure balanced individual accountability with collective reasoning, fostering introspective engagement and dialogic learning.

(3) Reflective Reporting and Peer Evaluation

After each quiz session, quiz-giving teams submitted a 500-word reflective report via Google Docs, with each member contributing a minimum of 100 words. Reports could be written in either English or Chinese, as long as the language remained consistent throughout. These reflections addressed key aspects such as individual learning experiences, team dynamics, and insights into assessment design.

To promote transparency and mutual accountability, peer contribution scores (on a 1–5 scale) were submitted through FlipClass Assignments and made visible to all team members.

Students were provided with structured guidelines for reflective report writing, including (a) reflection on the quiz design process and learning experiences, (b) analysis of group dynamics and collaborative decision-making, (c) insights into assessment design principles and item validity, (d) evaluation of peer responses and feedback received, and (e) identification of areas for improvement in future quiz cycles. These guidelines ensured that reflections addressed both cognitive and metacognitive dimensions of the SGA experience while maintaining individual voice and authenticity. While these themes provided structure, students were explicitly encouraged to include additional reflections, personal insights, and unexpected discoveries that emerged during the collaborative process, ensuring that reports captured both guided reflection and authentic learner voice.

(4) Survey-Based Perception Measures

Two Likert-scale survey items were extracted from broader survey instruments administered as part of the comprehensive TQB-FC evaluation (detailed in the companion paper). These specific items were selected to capture affective responses most relevant to the SGA cycle: (a) “How engaged do you feel with the assessment design process in this course?” (6-point scale: 1=Not engaged at all, 6=Extremely engaged) and (b) “How satisfied are you with the quality of quizzes created by your peers?” (5-point scale: 1=Very unsatisfied, 5=Very satisfied). The broader surveys contained additional items related to other aspects of the flipped classroom model, but these two items were most pertinent to the present study’s focus on SGA.

(5) Open-Ended Conception Survey

To complement the quantitative data, an open-ended survey item invited students to articulate their evolving conceptions of assessment design. This item was administered at the end of both the fall and spring semesters. Student responses were qualitatively analyzed to trace shifts in their understanding of assessment as a learner-centered and socially mediated practice, with attention to how they perceived its role in fostering agency, reflection, and collaborative engagement.

(6) Learner Reflections Through Semi-Structured Interviews

To complement survey-based and artifact-driven data, two rounds of semi-structured interviews were conducted to investigate learners’ evolving perceptions of grammar instruction, learner autonomy, and assessment design within the TQB-FC framework. These interviews focused specifically on the HL phase, where student-generated assessments were most salient and learner agency was actively exercised.

A total of 15 students participated in two academic semesters. Participants were selected through stratified sampling based on final term grades to ensure representation across performance levels. Interviews were conducted in either English or Mandarin, depending on participant preference, and lasted approximately 30~45 minutes. All interviews were audio-recorded, transcribed, and anonymized prior to analysis.

Interview protocols varied across semesters. In the fall, Interview A explored four thematic domains: motivation to engage, instructional engagement, learner autonomy, and perceived challenges. Prompts invited students to reflect on their initial exposure to the flipped classroom structure, the role of collaborative quiz design, and the extent to which course tools supported independent learning and ownership. These insights directly informed instructional refinements for the spring semester, particularly in scaffolding collaborative tasks and enhancing feedback mechanisms.

In the spring, Interview B examined perception shifts, strategic engagement, cognitive autonomy, and feedback reflection. Students were asked to compare their experiences across semesters, describe changes in their learning strategies, and articulate how peer collaboration and feedback influenced their approach to quiz design and problem-solving. Responses illuminated learners' development as co-constructors of assessment and their growing metacognitive awareness. A summary of thematic domains is provided in Table 1.

Table 1

Summary of Interview Protocols: Purpose and thematic focus of semi-structured interviews across Fall and Spring cohorts

Semester	Purpose	Thematic Domains
Fall (Interview A)	Explore perceptions of grammar instruction, autonomy, engagement, and challenges	Motivation to Engage, Instructional Engagement, Learner Autonomy, Perceived Challenges
Spring (Interview B)	Examine learner growth following refinements to the TQB-FC model	Perception Shifts, Strategic Engagement, Cognitive Autonomy, Feedback Reflection

5. Data Analysis Procedures

(1) Cognitive Complexity Coding

A corpus of 685 student-generated quiz items was analyzed to assess cognitive complexity using a simplified Bloom-based rubric adapted from Bloom's Revised Taxonomy. To enhance coding clarity and inter-rater reliability, the original six levels (*Remember, Understand, Apply, Analyze, Evaluate, and Create*) were consolidated into four operational categories: Foundational (Remember + Understand), Procedural (Apply), Analytical (Analyze + Evaluate), and Generative (Create). This structure was tailored to grammar-focused tasks and informed by prior research on cognitive classification in flipped learning environments (e.g., Gebregziabher et al., 2025).

Two trained raters independently coded all items, achieving inter-rater reliability of $\kappa = 0.84$, indicating near-perfect agreement. Discrepancies were resolved through discussion and rubric refinement. This coding enabled the identification of developmental shifts in learners' cognitive engagement—particularly their increasing ability to design tasks that moved beyond surface-level recall toward higher-order reasoning, such as error analysis, justification, and sentence reconstruction.

(2) Thematic Analysis

Interview transcripts, collaborative reports, and open-ended survey responses were analyzed inductively to identify emergent themes related to assessment literacy, metacognitive development, learner agency, and

instructional responsiveness. Coding was guided by Fulcher's (2012) assessment literacy framework, which emphasizes learners' understanding of assessment purposes, design, and feedback processes. This framework was selected for its alignment with the study's emphasis on student-generated assessment and learner-centered pedagogy.

Thematic coding was refined through iterative memoing to ensure theoretical coherence and contextual depth. This approach allowed for the emergence of nuanced patterns in how learners engaged with grammar content, navigated collaborative assessment design, and reflected on their evolving roles as co-constructors of learning. Thematic categories were continuously reviewed and adjusted to capture developmental shifts in learner cognition and pedagogical responsiveness across the intervention.

(3) Descriptive Statistics

Facility indices and peer-evaluation ratings were analyzed using IBM SPSS Statistics (Version 28.0) to track changes in item difficulty and learner satisfaction across cycles. Median and IQR values were reported to account for non-normal distributions.

(4) Pre-/Post-Test Comparison

Grammar proficiency was measured using instructor-designed tests aligned with course objectives. Pre- and post-test scores were analyzed in IBM SPSS Statistics (Version 28.0) to identify gains, which were interpreted as indirect indicators of improved assessment literacy and deeper grammatical understanding.

(5) Treatment of Likert-Scale Data

Following recent conventions in educational research (Carifio & Perla, 2008; Sullivan & Artino, 2013), Likert-scale responses were treated as interval-level data for primary analysis, with parametric tests (paired *t*-tests, Cohen's *d*) serving as the main statistical approach. This decision was based on the practical utility of these measures in educational contexts and their widespread interpretability. Although Likert-scale responses are technically ordinal, this treatment allowed for meaningful descriptive analysis using means and standard deviations to summarize central tendencies, while medians were also included to reflect distributional characteristics. To ensure statistical robustness and acknowledge the ordinal nature of the underlying scale, non-parametric tests (Wilcoxon signed-rank tests, rank-biserial correlation) were conducted as supplementary analyses to validate findings. Effect sizes were calculated using both Cohen's *d* and rank-biserial *r,* and both parametric and non-parametric approaches yielded consistent conclusions, strengthening confidence in the results. This dual approach balances interpretive clarity with methodological rigor while leveraging the practical utility of interval-level analysis in longitudinal educational research.

The different scale formats were chosen to reflect the distinct nature of each construct and optimize response quality. The 6-point scale for engagement was selected to eliminate the neutral midpoint option, encouraging respondents to lean toward either positive or negative engagement—appropriate for measuring internal motivational dispositions where neutrality is less meaningful. The 5-point scale for satisfaction included a neutral option to accommodate genuine ambivalence about peer-generated materials, recognizing that satisfaction represents an evaluative judgment where neutrality can be a valid response. This differentiation acknowledged that engagement (internal disposition) and satisfaction (evaluative judgment) represent conceptually different constructs that benefit from distinct measurement approaches.

IV. Results

This section presents findings from the implementation of the SGA cycle within the TQB-FC framework.

Data sources included student-generated quiz items, grammar proficiency tests, reflective reports, survey responses, and post-semester interviews. Results are organized into five domains: (1) cognitive-complexity evolution, (2) grammar proficiency gains, (3) assessment-literacy development, (4) learner satisfaction, and (5) qualitative insights from interviews.

1. Cognitive Evolution Through Scaffolded Quiz Design

Analysis of 685 student-authored quiz items revealed a marked progression in cognitive engagement across instructional cycles. As shown in Table 2, by the final semester, 89.6% of tasks aligned with Procedural, Analytical, or Generative levels—a substantial increase from 66.4% in the initial cycle. This upward shift reflects learners' growing capacity to operationalize grammatical knowledge within context-rich, communicative formats.

A chi-square test confirmed a significant redistribution of items across cognitive levels, $\chi^2(3, N = 685) = 162.4$, $p < .001$, Cramer's $V = .49$, indicating a large effect size. The most substantial gain occurred in the Procedural category (+14.5%), reflecting learners' increased ability to apply grammatical knowledge in contextually meaningful ways—such as sentence rewriting and transformation. This rise suggests that scaffolded quiz design effectively supported movement from passive recognition toward active manipulation of form and function.

Additional gains were observed in the Analytical category (+2.7%) and the emergence of Generative-level tasks (+6.1%), particularly in open-ended sentence reconstruction. These shifts signal a conceptual reframing of grammar—from static form recognition to dynamic meaning-making and reasoning—consistent with Bloom's Revised Taxonomy (Anderson & Krathwohl, 2001) and sociocultural perspectives on learner agency and mediated cognition (Vygotsky, 1978). Together, the upward trajectory across all non-foundational levels underscores the pedagogical impact of iterative, student-centered assessment design.

Table 2

Cognitive Level Distribution by Semester (Simplified Bloom-Based Rubric)

Simplified Level	Fall *n* (%)	Spring *n* (%)	Δ%	Sample Formats
Foundational (Remember+Understand)	101 (33.6)	40 (10.4)	–23.2	Comparative/superlative fill-in; Cloze conjunction MCQ
Procedural (Apply)	138 (45.9)	233 (60.4)	+14.5	Sentence rewriting
Analytical (Analyze + Evaluate)	61 (20.5)	89 (23.2)	+2.7	Error-analysis justification; Present perfect analysis
Generative (Create)	0 (0)	23 (6.1)	+6.1	Sentence reconstruction

Importantly, the grammatical content covered across semesters influenced cognitive demand. Foundational units in the Fall (e.g., simple present, pronouns, quantity expressions) were more frequently associated with Foundational and Procedural-level tasks. In contrast, Spring units (e.g., modals, adjective clauses, passive voice) elicited a higher proportion of Analytical and Generative-level items. This pattern suggests that linguistic complexity and instructional sequencing shaped the depth of learner-generated assessments.

To illustrate this relationship, Table 3 presents a taxonomy–grammar matrix mapping the simplified Bloom-based rubric against grammar units taught in each semester. The matrix highlights dominant cognitive levels per unit and includes representative task formats. It reinforces the finding that grammatical structure mediates cognitive complexity, with certain forms (e.g., modals, embedded clauses) more conducive to higher-order reasoning. Future research may further explore how grammar-task interactions foster metacognitive engagement and inform instructional design.

Table 3

Taxonomy–Grammar Matrix

Grammar Unit	Dominant Cognitive Level(s)	Sample Task Type
Fall Semester (Units 1–7)		
Simple present & frequency words	Foundational → Procedural	Cloze, rewriting
Present continuous & future forms	Foundational → Procedural	Sentence rewriting
Simple past & habitual past ("used to")	Procedural	Contextual rewriting
Possessive forms, pronouns, question formation	Foundational → Procedural	MCQ, sentence correction
Nouns, There + Be, quantity expressions	Foundational	Cloze, fill-in
Modifiers & adverbs	Procedural	Sentence transformation
Time expressions & past continuous	Procedural	Error identification
Spring Semester (Units 8–14)		
Modals	Analytical → Generative	Justification, reconstruction
Present perfect & continuous	Analytical	Tense comparison, error analysis
Gerunds & infinitives	Procedural → Analytical	Sentence rewriting, distractor design
Adjective clauses	Analytical → Generative	Sentence synthesis
Comparatives & superlatives	Foundational → Procedural	Fill-in, rewriting
Passive & active voice	Analytical	Transformation, peer critique
Articles, other/another, indefinite pronouns	Procedural	Cloze, distractor revision

2. Grammar Proficiency Gains

Students demonstrated meaningful and statistically significant growth in their ability to apply grammar in context across both semesters. For the Fall cohort (n = 61 pre-test, 55 post-test), average scores rose from 64.8 to 79.5, marked by a large effect size (Z = –5.87, p < .000001, d = 0.85). Spring learners (n = 55) mirrored this progress, with scores climbing from 65.4 to 82.7 (Z = –6.15, p < .000001, d = 0.90) (see Table 4).

Table 4

Grammar Proficiency Gains Across Semesters

Cohort	Pre-Test *Mean*	Post-Test *Mean*	*Z*-Score	*p*-value	Effect Size *(d)*
Fall	64.8	79.5	–5.87	< .000001	0.85
Spring	65.4	82.7	–6.15	< .000001	0.90

No initial proficiency gaps were detected between continuing and transfer students (Fall: p = .37), supporting the integrity of observed outcomes. Learners engaged deeply with the flipped model—surpassing meta-analytic benchmarks for grammar instruction (Baig & Yadegaridehkordi, 2023; Norris & Ortega, 2000). Interview data suggest metacognitive transfer: *"Designing quiz items made me double-check rules when I faced similar patterns in the post-test."* Ultimately, the TQB-FC framework fostered ownership of learning, turning grammar from rule-based memorization into a shared language for expression and connection.

3. Assessment-Literacy Development

Three interrelated themes emerged from 29 reflective reports and 92 open-ended survey responses, indicating growth in learners' assessment literacy:

(1) Construct Alignment Awareness

By Spring, 86% of reports explicitly connected item purpose to lesson outcomes. Learners demonstrated an ability to articulate the pedagogical intent behind their quiz items: *"Question 3 targets the sub-rule on inversion after 'scarcely,' which half our cohort missed in Quiz 11."*

(2) Evidence-Based Reasoning

Students increasingly referenced facility indices and revised distractors accordingly: *"0.41 suggests moderate difficulty, so we kept distractor B but changed C to make it less obvious."*

(3) Metacognitive Reflection

Learners adopted a "teacher lens," critically evaluating item plausibility and linguistic precision: *"I realized my distractor C was implausible because it violated subject–verb agreement."*

These themes reflect a shift from passive test-taking to active assessment design, aligning with Fulcher's (2012) framework and reinforcing the role of SGA in cultivating reflective, data-informed learners.

4. Learner Satisfaction and Perception Shifts

Survey data revealed a significant increase in satisfaction with peer-generated quizzes across semesters. Mean scores rose from $M = 3.17$ ($SD = 0.82$) in Fall ($n = 55$) to $M = 4.05$ ($SD = 0.70$) in Spring ($n = 48$), $t(47) = –6.42$, $p < .001$, $d = 1.21$, indicating a large effect size. Given the ordinal nature of Likert-scale data, a Wilcoxon signed-rank test was also conducted to validate these findings. Results confirmed a statistically significant shift in satisfaction ratings ($Z = –3.94$, $p < .00009$ in Fall; $Z = –4.28$, $p < .00004$ in Spring), with moderate-to-large effect sizes ($r = .53$ and .58, respectively) (see Table 5). These outcomes reinforce the robustness of the observed perceptual gains.

Table 5

Descriptive and inferential statistics for satisfaction with peer-generated quizzes, by semester

Statistic	Fall	Spring
Sample Size	55	48
Satisfaction Rate (%)	80.0%	83.3%
Median Rating	4	4
Wilcoxon Z	–3.94	–4.28
***p*-Value**	< .00009	< .00004
Effect Size (*r*)	0.53	0.58
95% CI Lower	0.41	0.45
95% CI Upper	0.66	0.70

Note. The Wilcoxon signed-rank test was used for within-semester comparison of satisfaction ratings. Effect size was calculated using rank-biserial correlation (r) and interpreted using Cohen's guidelines. All comparisons were statistically significant at $p < .0001$. In Spring, satisfaction ratings were positively associated with post-test grammar scores ($r = .44$, $p = .008$), suggesting a link between perceived task value and learning outcomes.

Qualitative comments reflected a perceptual shift:

(1) Early skepticism: *"Peers may trick us or ask confusing questions."*

(2) Later endorsement: *"Classmates' quizzes pinpointed my weak spots better than textbook drills."*

Students reported increased trust in peer-authored materials and recognized the diagnostic value of collaboratively

designed assessments. The visibility of peer evaluation scores and structured feedback mechanisms reinforced transparency and accountability. These findings suggest that satisfaction was not merely a function of enjoyment but a reflection of perceived efficacy and alignment with learning goals.

5. Interview Insights: Depth, Format, and Challenge

Post-semester interviews with 15 participants provided nuanced insights into learner experiences. Three themes emerged:

(1) Design as Deep Engagement: *"Even though I gave a quiz once, I spent weeks refining items. It made me think like a teacher."*

(2) Format Preferences and Cognitive Demand: *"For the Analyze level, we preferred error identification instead of multiple-choice; it triggered discussion."*

(3) Creative Tension and Peer Critique: *"I wanted to make a clever distractor, but my teammates said it was too confusing." "Peer critique was helpful but stressful. I didn't want to disappoint my group."*

These insights highlight the emotional and cognitive dimensions of learner-centered assessment and underscore the importance of scaffolding not only linguistic content but also collaborative processes.

To further illustrate the qualitative patterns observed across interviews, reflective reports, and open-ended survey responses, Table 6 summarizes the emergent themes from the thematic analysis.

Table 6

Emergent Themes from Thematic Analysis of Collaborative Reports, Survey Responses, and Interviews

Theme	Description	Illustrative Focus	Sample Quote
Assessment Literacy	Understanding of assessment principles, item validity, and cognitive demand	Reflections on quiz design, error analysis, and peer review	"We realized our question was too easy—next time we'll add a distractor to test deeper thinking."
Metacognitive Development	Awareness of learning strategies, self-monitoring, and reflective reasoning	Strategic preparation, report writing, feedback uptake	"After seeing our mistakes, I started checking grammar rules before writing quiz items."
Learner Agency	Ownership of learning, decision-making in task design, and peer negotiation	Quiz authorship, team consensus-building, autonomy claims	"Our group argued about the correct answer, but it helped me see grammar from different angles."
Instructional Responsiveness	Perception of instructional scaffolding, clarity, and adaptability	Feedback from instructor/TA, tool usability, task clarity	"The teacher didn't fix our flawed item—so we had to explain it ourselves. That made us think harder."

6. Conceptions of Assessment: Qualitative Shifts from Open-Ended Survey

To complement the quantitative and reflective data, an open-ended survey item was administered at the end of both semesters, inviting students to articulate their evolving conceptions of assessment design. Responses (n = 92) were thematically analyzed to trace shifts in learners' understanding of assessment as a learner-centered and socially mediated practice.

Three core themes emerged:

(1) Learner Agency: Students increasingly viewed assessment as a tool for self-regulation and growth.

"Assessment should help me see what I understand, not just what I got wrong."

(2) Collaborative Engagement: Responses highlighted the value of peer interaction and shared responsibility. *"I liked when we discussed answers—it made me think more deeply."*

(3) Reflective Practice: Learners recognized assessment as a meta-cognitive process. *"Posting answers and checking them helped me realize where I need to improve."*

Fall semester responses tended to emphasize clarity and correctness, while Spring responses reflected deeper engagement with assessment as a dialogic and formative process. This progression aligns with the sociocultural underpinnings of the TQB-FC framework, suggesting that repeated exposure to scaffolded, student-generated tasks fosters more nuanced and empowered conceptions of assessment.

7. Summary of Findings

The integration of SGA within the TQB-FC framework yielded statistically and pedagogically significant outcomes. Learners demonstrated increased cognitive complexity in quiz design, improved grammar proficiency, and enhanced assessment literacy. Satisfaction with peer-generated materials rose markedly, and qualitative data revealed deep engagement with both linguistic form and pedagogical function. The recursive structure of the SGA cycle—design, administer, reflect—supported both meta-cognitive growth and collaborative accountability, positioning learners as active agents in their own assessment journeys.

V. Discussion

This study explored how SGA, embedded within a TQB-FC framework, can foster grammar learning, cognitive complexity, and assessment literacy in a Taiwanese EFL context. The findings suggest that when learners are positioned not merely as recipients of instruction but as co-constructors of assessment, their engagement with grammar deepens—both cognitively and meta-cognitively.

1. Reframing Grammar as Meaning-Making

This study revealed a notable shift in learners' conceptualization of grammar—from a static, rule-based system to a dynamic, communicative resource. This reframing was evident in both the increasing complexity of quiz items and the reflective discourse in student reports. The emergence and expansion of Apply through Create-level tasks (89.6% by the final cycle) marks a pedagogical departure from traditional grammar drills toward cognitively demanding, context-rich challenges. In particular, the surge in Procedural-level engagement (+14.5%) suggests that learners increasingly viewed grammar as a tool for constructing meaning rather than merely recalling rules. This evolution aligns with sociocultural perspectives on language learning, where cognition is mediated through purposeful activity and scaffolded interaction (Vygotsky, 1978).

This cognitive transformation also resonates with Schmidt's Noticing Hypothesis (2001), which posits that conscious attention to linguistic form is essential for acquisition. In the present study, noticing was not incidental—it was deliberately cultivated through quiz design, peer negotiation, and error analysis. These activities positioned learners as active interrogators of grammar, prompting them to repurpose structures and embed them in authentic communicative contexts. Such engagement echoes Lu's (2025) findings that grammar-based question creation using AI corpora enhances both grammatical accuracy and writing fluency.

Moreover, this process reflects the metacognitive activation described by Panadero et al. (2017), who argue that student-generated tasks stimulate planning, monitoring, and evaluative reasoning. In this study, learners didn't merely observe grammar—they constructed it, critiqued it, and refined it collaboratively. The recursive nature of

the SGA cycle thus fostered a shift from passive reception to strategic authorship, reinforcing grammar as a vehicle for meaning-making and reflective agency.

2. Learners as Designers: Agency and Accountability

This transformation from passive recipients to active designers reflects Vygotsky's concept of mediated learning, where tools (in this case, assessment design protocols) reshape cognitive processes. Students' adoption of a "teacher lens" demonstrates what Fulcher (2012) terms "critical assessment literacy"—the ability to interrogate assessment purposes, validity, and consequences.

The scaffolded nature of the TQB-FC framework was crucial in supporting this transition. Rather than overwhelming students with complete design autonomy, the three-phase structure (SL, PL, HL) provided graduated responsibility. Students first engaged with instructor-designed materials, then participated in peer evaluation, and finally took ownership of item creation. This progression mirrors Bloom's taxonomy in reverse—moving from evaluation and analysis to creation and synthesis.

Furthermore, the collaborative dimension of quiz design fostered what Meijer et al. (2020) describe as "shared regulation"—where learning responsibility is distributed across team members. Students negotiated not only linguistic content but also pedagogical decisions: Which distractors were plausible? How could items target specific learning outcomes? What level of difficulty was appropriate for their peers? These negotiations required both subject-matter expertise and instructional sensitivity, positioning students as emerging teacher-researchers rather than mere content consumers.

3. Cognitive Complexity Through Collaborative Construction

The evolution of student-authored items illustrates the power of collaborative construction in fostering cognitive complexity. Initial quizzes mirrored textbook formats, privileging recall. Over time, guided by Bloom's-based modeling and iterative feedback, learners produced tasks demanding analysis, synthesis, and contextual interpretation—validating Anderson and Krathwohl's (2001) taxonomy as a lens for tracking cognitive depth.

This progression aligns with Vygotsky's Zone of Proximal Development (1978), where learners, supported by peers and instructors, operate just beyond their current capabilities. Stratified team formation and dual-phase quiz administration created fertile ground for dialogic learning. Gebregziabher et al. (2025) similarly found that flipped environments scaffold higher-order grammar reasoning, especially when students create and evaluate peer-generated items.

The decline in Understand-level items and the disappearance of Evaluate-level tasks suggest a gravitation toward formats that felt linguistically authentic and pedagogically manageable. Error identification and rewriting tasks emerged as preferred formats, balancing depth with clarity. Hernandez Sibo et al.'s (2024) synthesis on cognitive load in creative thinking supports this trend, highlighting how task design influences learners' ability to balance novelty and usefulness.

Importantly, the grammatical focus of each quiz cycle also shaped cognitive demand. Tasks involving tense sequencing, modals, or adjective clauses—introduced in later units—tended to elicit higher-order reasoning (e.g., Analyze, Create), while quizzes centered on quantity expressions or subject–verb agreement often remained at the Understand or Apply levels. This pattern suggests that linguistic content mediates cognitive complexity, with certain grammatical domains more conducive to abstract reasoning and syntactic manipulation. Table 3 provides a comprehensive mapping of this relationship, showing how different grammar units across both semesters correspond to varying cognitive demands and task formats.

This interplay between grammar content and cognitive level warrants deeper exploration, particularly in contexts where learners generate assessments collaboratively. Future studies might examine how specific

linguistic features interact with task type to shape metacognitive engagement and learner agency.

4. Metacognitive Gains Through Grammar Assessment

The statistically significant gains in grammar proficiency (Fall: $d = 0.85$; Spring: $d = 0.90$) reflect not only instructional effectiveness but also metacognitive transfer. Learners reported applying quiz-design strategies to their own test-taking, indicating that constructing assessments enhanced their ability to decode and apply grammatical rules in novel contexts.

This strategic transfer aligns with Brandmo et al.'s (2020) assertion that formative assessment can activate self-regulated learning when it supports the full cycle of forethought, performance, and reflection. Their model highlights how assessment tasks—particularly those embedded in collaborative contexts—foster co-regulation and socially shared regulation, enabling learners to negotiate meaning, monitor understanding, and refine strategies through peer interaction. In the present study, the recursive nature of the SGA cycle provided repeated opportunities for such regulation, reinforcing learners' strategic autonomy and deepening their engagement with grammatical form and function.

This outcome also supports Swain's Output Hypothesis (2005), which emphasizes the role of production in noticing and internalizing linguistic gaps. In this study, output extended beyond speaking and writing to include assessment artifact design—requiring learners to externalize grammatical understanding and refine it through peer critique. Molina et al. (2023) similarly argue that item design fosters integrative complexity, enabling learners to differentiate and integrate knowledge in cognitively demanding ways.

The absence of initial proficiency gaps between continuing and transfer students further validates the accessibility of the model. Regardless of background, learners engaged meaningfully with the flipped framework, surpassing meta-analytic benchmarks for grammar instruction (Baig & Yadegaridehkordi, 2023; Norris & Ortega, 2000).

5. Emotional Dimensions of Peer Assessment

While cognitive and linguistic gains were evident, the emotional terrain of peer assessment proved more complex. Interviews revealed both appreciation and anxiety—learners valued the richness of the design process but struggled with critique and role negotiation. This tension underscores the need for humanized scaffolding, where emotional safety is cultivated alongside academic rigor.

The visibility of peer evaluation scores and collaborative platforms like Google Docs helped mitigate these challenges by reinforcing transparency and shared ownership. Still, Mphahlele's (2024) review of collaborative assessment in open-distance learning contexts cautions that group dynamics and technology access can complicate engagement. Future iterations may benefit from structured peer feedback protocols and reflective prompts that explicitly address emotional dynamics within teams.

As summarized in Table 4 (Results, Section 5), thematic analysis revealed four interrelated dimensions of learner experience: assessment literacy, metacognitive development, learner agency, and instructional responsiveness. These themes reflect not only cognitive and pedagogical growth but also the emotional and interpersonal complexity of student-generated assessment. Sample quotes illustrate how learners negotiated task demands, internalized instructional goals, and developed a "teacher lens" through collaborative design and critique.

6. Grammar Complexity as Cognitive Catalyst

The relationship between grammatical content and cognitive complexity revealed systematic patterns that merit theoretical consideration. Units involving modal verbs, adjective clauses, and voice distinctions consistently

elicited higher-order thinking tasks, while topics such as quantity expressions and basic tense forms remained predominantly at foundational levels.

This pattern reflects what Bulté and Housen (2012) term "inherent complexity" versus "relative complexity." Modals require learners to manipulate meaning relationships (possibility, necessity, obligation), naturally demanding analysis and evaluation. Similarly, adjective clause construction involves syntactic embedding and referential relationships that challenge learners to create and synthesize. In contrast, subject-verb agreement or article usage, while essential, involves more rule-based applications that lend themselves to recognition and procedural practice.

From a cognitive perspective, these complex grammatical structures act as "thinking tools" (Vygotsky, 1978) that mediate higher-order reasoning. When students design tasks around modals or embedded clauses, they must consider not just form but also function, context, and communicative intent. This cognitive load naturally elevates their engagement from memorization toward analysis and creation.

These findings have important implications for curriculum sequencing and task design. Instructors seeking to promote higher-order thinking through SGA should consider introducing complex grammatical structures earlier in the semester, providing scaffolding to help students navigate both linguistic and cognitive demands. Conversely, foundational structures might be paired with more cognitively challenging task formats (error analysis, peer teaching) to maintain intellectual engagement.

7. Limitations

While the findings offer valuable insights into the pedagogical potential of SGA within a TQB-FC framework, several limitations warrant consideration.

(1) Contextual Specificity

The study was conducted within a single Taiwanese technical university, which may limit the generalizability of findings to other institutional types or cultural contexts. Learner responses and engagement patterns may differ in secondary education settings, liberal arts institutions, or non-Asian EFL environments.

(2) Assessment Format Sensitivity

Although Bloom's Revised Taxonomy was introduced and sample quizzes were provided, students retained autonomy over question formats. Most opted for discrete-point grammar tasks (e.g., multiple choice, error correction), which may reflect both cognitive comfort and perceived clarity. While this learner-driven preference supports agency, it may have limited exploration of open-ended or discourse-level formats that demand broader linguistic production and integrative reasoning.

In addition to format preferences, the grammatical content itself appeared to influence cognitive demand. Tasks involving modals, conditional structures, or clause embedding—typically introduced in later units—were more likely to elicit higher-order reasoning, while quizzes focused on article usage, quantity expressions, or subject-verb agreement tended to remain at the Understand or Apply levels. This suggests that linguistic complexity and instructional sequencing may shape not only the form but also the depth of learner-generated assessments. Future studies could examine how specific grammatical domains interact with task type to mediate metacognitive engagement and cognitive load.

(3) Self-Report and Reflective Bias

Learner perceptions were captured through reflective reports and survey instruments, which are inherently subjective. While triangulated with performance data, these sources may be influenced by social desirability. retrospective bias, or limited metacognitive awareness—particularly among lower-proficiency learners.

(4) Team Dynamics and Peer Influence

The collaborative nature of the SGA cycle introduced variability in team dynamics, which may have affected task quality and learner experience. While stratified grouping aimed to balance proficiency levels, individual contributions and peer scaffolding were not systematically tracked, making it difficult to isolate the effects of collaboration from individual cognitive growth.

Despite these limitations, several design features of the TQB-FC framework suggest potential transferability beyond the specific study context. The collaborative assessment design principles underlying SGA draw on universal pedagogical constructs, such as peer scaffolding, metacognitive reflection, and authentic task engagement, making them broadly applicable across diverse educational settings.

The framework's modular structure (self-learning, peer learning, hands-on learning) offers flexibility for adaptation to different institutional contexts by adjusting technology requirements, group sizes, and scaffolding intensity. The simplified Bloom-based rubric proved accessible to students with diverse proficiency levels, suggesting particular applicability across Asian EFL contexts where exam-oriented and communicative pedagogies often coexist.

However, certain contextual factors may influence implementation effectiveness. In collectivist educational cultures common across East Asia, the peer evaluation and group accountability mechanisms may resonate particularly strongly, as evidenced by the study's high satisfaction rates and commitment to group consensus-building. The emphasis on collaborative reflection and peer accountability aligns with collective educational values prevalent in these settings. Western individualistic contexts might require modified scaffolding approaches to achieve similar levels of collaborative engagement.

The model's emphasis on gradual responsibility transfer (SL→PL→HL phases) appears particularly suitable for educational contexts where students have limited prior experience with learner-centered pedagogy, suggesting potential applicability in secondary education settings and universities transitioning from traditional teacher-centered approaches. Technical universities and vocational institutions may find the model especially relevant due to its emphasis on practical skill application and collaborative problem-solving—competencies directly transferable to the professional contexts these institutions serve.

Future research should systematically examine how the model performs across different institutional types (secondary schools, liberal arts colleges, vocational programs) and cultural settings to establish broader generalizability and identify necessary adaptations for optimal implementation in diverse contexts.

VI. Conclusion

This study examined the pedagogical impact of SGA embedded within a TQB-FC framework in a Taiwanese EFL context. The intervention yielded significant gains in grammar proficiency, cognitive complexity, and assessment literacy, demonstrating that when learners are positioned as co-constructors of assessment, grammar instruction becomes a participatory, reflective, and intellectually rigorous endeavor.

1. Implications for Curriculum and Instruction

The findings underscore the need to move beyond procedural grammar instruction toward learner-centered, cognitively rich engagement. The emergence of Apply through Create-level tasks and the decline of surface-level formats suggest that students, when scaffolded appropriately, are capable of designing and navigating complex linguistic challenges.

Curriculum designers should consider embedding SGA cycles into existing grammar modules, integrating

quiz design, peer evaluation, and reflective reporting as core components. Simplified Bloom-based rubrics and collaborative platforms can support this integration, ensuring that cognitive complexity and learner agency are systematically cultivated. The model's accessibility across learner profiles—including part-time workers and transfer students—further affirms its scalability across diverse educational settings.

For educators, the study invites a reimagining of the teacher's role—from sole assessor to facilitator of assessment literacy and cognitive growth. Professional development programs should equip teachers to scaffold student-designed tasks, model cognitive complexity, and navigate the emotional dimensions of peer assessment. Training might include:

(1) Designing Bloom-aligned grammar tasks collaboratively with students.

(2) Facilitating peer feedback protocols that balance critique with emotional safety.

(3) Using student-generated artifacts as diagnostic tools for instructional responsiveness.

(4) Cultivating a classroom culture of shared ownership and reflective inquiry.

Such practices not only enhance instructional effectiveness but also foster inclusive, dialogic learning environments where students interrogate, apply, and extend their grammatical knowledge.

2. Future Research Directions

Building on these findings, future research should explore:

(1) How learners transfer quiz-design strategies across genres, modalities, and linguistic domains.

(2) The long-term impact of SGA cycles on grammar retention, assessment literacy, and learner agency.

(3) Format sensitivity in student-generated tasks, particularly the cognitive and emotional affordances of different item types.

(4) Effective models of teacher professional development that support participatory assessment practices.

By continuing to investigate these dimensions, researchers and practitioners can refine sustainable, inclusive models of grammar instruction that empower learners not just to perform but to understand, critique, and co-create the very tools of their learning.

References

Anderson, L.W., & Krathwohl, D.R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Longman.

Baig, M.I., & Yadegaridehkordi, E. (2023). Flipped classroom in higher education: A systematic literature review and research challenges. *International Journal of Educational Technology in Higher Education*, *20*(1), 61. https://doi.org/10.1186/s41239-023-00430-5

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Brandmo, C., Panadero, E., & Hopfenbeck, T.N. (2020). Bridging classroom assessment and self-regulated learning. *Assessment in Education: Principles, Policy & Practice*, *27*(4), 319–331. https://doi.org/10.1080/0969594X.2020.1803589

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in*

SLA (pp. 21–46). John Benjamins.

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, *42*(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Elbaum, S.N. (2016). *Grammar in Context 2* (6th ed.). Heinle Cengage Learning.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041

Gebregziabher, H.A., Filate, A.Y., & Bishaw, K.S. (2025). Grammar learning in a flipped classroom: Measuring achievement and students' perceptions. *Discover Education, 4*, 265. https://doi.org/10.1007/s44217-025-00641-0

Hernandez Sibo, I.P., Gomez Celis, D.A., & Liou, S. (2024). Exploring the landscape of cognitive load in creative thinking: A systematic literature review. *Educational Psychology Review, 36*, Article 24. https://doi.org/10.1007/s10648-024-09866-1

Hung, S., & Huang, H. (2019). Standardized proficiency tests in a campus-wide English curriculum: A washback study. *Language Testing in Asia, 9*(1), 1–17. https://doi.org/10.1186/s40468-019-0096-5

IBM Corp. (2021). *IBM SPSS Statistics for Windows* (Version 28.0) [Computer software]. IBM Corp.

Katz, L., Carlgren, D., Wright-Maley, C., Hallam, M., Forder, J., Milner, D., & Finestone, L. (2024). Student-generated multiple-choice questions: A Java and web-based tool for students to create multiple choice tests. *The Canadian Journal for the Scholarship of Teaching and Learning, 15*(2). https://doi.org/10.5206/cjsotlrcacea.2024.2.16625

Li, M., & Wilson, J. (2025). AI-integrated scaffolding to enhance agency and creativity in K-12 English language learners: A systematic review. *Information, 16*(7), 519. https://doi.org/10.3390/info16070519

Lin, S.W. (2016). The power of General English Proficiency Test on Taiwanese society and its tertiary English education. In G. Yu & Y. Jin (Eds.), *Assessing Chinese learners of English* (pp. 267–284). Palgrave Macmillan. https://doi.org/10.1057/9781137449788_13

Liu, C.C., Hwang, G.J., Yu, P., Tu, Y.F., & Wang, Y. (2025). Effects of an automated corrective feedback-based peer assessment approach on students' learning achievement, motivation, and self-regulated learning conceptions in foreign language pronunciation. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-025-10484-z

Lu, C. (2025). AI-generated corpus learning and EFL learners' learning of grammatical structures, lexical bundles, and willingness to write. *PLOS ONE, 20*(7), e0321544. https://doi.org/10.1371/journal.pone.0321544

McKenney, S., & Reeves, T. (2018). *Conducting educational design research* (2nd ed.). Routledge. https://doi.org/10.4324/9781315105642

Meijer, H., Hoekstra, R., Brouwer, J., & Strijbos, J.W. (2020). Unfolding collaborative learning assessment literacy: A reflection on current assessment methods in higher education. *Assessment & Evaluation in Higher Education, 45*(8), 1222–1240. https://doi.org/10.1080/02602938.2020.1729696

Molina, I., Molina-Perez, E., Sobrino, F., Tellez-Rojas, M.A., Zamora-Maldonado, H.C., Plaza-Ferreira, M., Orozco, Y., Espinoza-Juarez, V., Serra-Barragán, L., & De Unanue, A. (2023). Current research trends on cognition, integrative complexity, and decision-making: A systematic literature review using activity theory and neuroscience. *Frontiers in Psychology, 14*, Article 1156696.

https://doi.org/10.3389/fpsyg.2023.1156696

Mphahlele, R. (2024). A review of research on collaborative assessments in the open distance and e-learning environment. *Journal of Learning for Development, 11*(2), 206–219. https://files.eric.ed.gov/fulltext/EJ1436999.pdf

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

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, Article 422. https://doi.org/10.3389/fpsyg.2017.00422

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and academic achievement: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Pastore, S. (2023). Assessment literacy in the higher education context: A systematic review. *Intersection: A Journal at the Intersection of Assessment and Learning, 4*(1), 1–24. https://files.eric.ed.gov/fulltext/EJ1386045.pdf

Pawlak, M. (2024). Grammar learning strategies: Towards a pedagogical intervention. *Language Teaching Research Quarterly, 39*, 174–191. https://doi.org/10.32038/ltrq.2024.39.12

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780.003

Sullivan, G.M., & Artino, A.R., Jr. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. https://doi.org/10.4300/JGME-5-4-18

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 495–508). Routledge. https://doi.org/10.4324/9781410612700-38.

Vygotsky, L.S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Xia, Q., Weng, X., Ouyang, F., Lin, T.J., & Chiu, T.K. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education, 21*(1), 1–22. https://doi.org/10.1186/s41239-024-00468-z

Yu, F.-Y., & Wu, W.-S. (2017). Student-generated feedback for online student-generated multiple-choice questions: Effects on question-generation performance and perspective-taking development. *International Conference on Computers in Education*. https://library.apsce.net/index.php/ICCE/article/view/2217

Zainuddin, Z., & Halili, S.H. (2016). Flipped classroom research and trends from different fields of study. *International Review of Research in Open and Distributed Learning, 17*(3), 313–340. https://doi.org/10.19173/irrodl.v17i3.2274