

Received: 27 May 2026; Revised: 28 July 2026; Accepted: 09 September 2026; Published: 17 September 2026

Archs Sci. (2026) Volume 76(2), Pages 72-89, Paper ID 2026208.
<https://doi.org/10.68304/as/76208>

ARCS-Informed Vocational English Writing: An Instructional and Assessment Design

Jie Shao

School of Culture, Tourism and International Education, Henan University of Industry Technology, Henan Nanyang, 473000, China

Corresponding author: Jie Shao (e-mail: jessica_shao958@163.com).

Abstract Motivational instructional design becomes evaluable when its activities, comparison condition, assessment responses, and intended interpretations are specified together. This design-and-methods article develops an ARCS-informed specification for vocational English writing, integrating attention, relevance, confidence, and satisfaction with communicative task design and constructive alignment. A purposive literature synthesis informs two 45-minute lessons, a content-matched comparison, fictional practice and assessment materials, and a four-dimension analytic rubric. The primary outcome is defined as the mean of two independent rater totals on a 0–16 scale. A logical examination shows that the composite permits compensation between dimensions, motivates separate factual-completeness checks, and identifies the assumptions under which assessment-form counterbalancing removes a fixed form difference. The prospective evaluation specifies class-level allocation, baseline and form adjustment, delayed near-transfer assessment, rater procedures, and implementation records. The contribution is a reproducible instructional specification and an explicit argument linking its proposed measurements to bounded educational claims. No classroom implementation, psychometric validation, or treatment effect is reported. The materials support formative use and empirical feasibility work; their evaluation concerns a defined family of customer-response tasks and does not establish general workplace proficiency. The article provides a concrete basis for examining an instructional package without equating motivational response, polished language, and accurate communication.

Index Terms vocational English; ARCS; motivational instructional design; educational assessment; constructive alignment; reproducible specification

1. Introduction

Vocational English writing requires learners to communicate a practical decision accurately to an intended recipient. In a customer inquiry, a grammatically fluent response can still fail if it promises an unavailable service, omits a confirmation deadline, or leaves the next action unclear. Conversely, a response may communicate the essential decision despite minor language errors. Instruction and assessment therefore need an explicit account of what a successful response accomplishes. The action-oriented perspective of the Common European Framework of Reference for Languages (CEFR) provides a communicative starting point, while recognizing that reception, production, interaction, and mediation are distinct activities [1].

The attention, relevance, confidence, and satisfaction (ARCS) model supplies an established approach to motivational instructional design [2]. Its educational applications have been reviewed by Li and Keller [3]. The present problem is consequently not how to introduce four familiar motivational categories into another subject. It is how to translate those categories into a sufficiently explicit lesson and comparison that a reader can determine what learners receive, what performance is assessed, and what a subsequent result would mean.

Constructive alignment addresses the relationship between intended performance, learning activities, and assessment [4]. Combining these perspectives is useful only if the resulting specification resolves practical decisions that the category names alone leave open.

Related language-education studies illustrate why the setting and endpoint matter. Sun and Asmawi examined WeChat-based business English writing instruction among Chinese university learners [5]. Research on teachers' pedagogical content knowledge in classroom-based assessment emphasizes the assessment practices through which classroom evidence is interpreted [6]. Liu and colleagues investigated relationships among motivation, informal digital English learning, and enjoyment [7]. These works concern different interventions, populations, and questions. They provide context for the current application, without establishing that an enjoyable activity or a particular delivery technology improves the constrained writing performance specified here.

A further difficulty concerns the meaning of transfer. Changing a room-booking inquiry to an equipment-booking inquiry changes the vocabulary and practical details while retaining much of the communicative structure. Such a task can test application beyond a memorized practice response, but it

cannot represent the full range of workplace communication. Barnett and Ceci show why transfer claims need to specify the dimensions across which learning is expected to extend [8]. The current design accordingly targets near transfer within a bounded family of written customer responses, with a delayed assessment of that same family. Spoken negotiation, unfamiliar genres, workplace behavior, and long-term occupational outcomes are outside its stated endpoint.

This design-and-methods article addresses three questions. First, how can ARCS-informed activities be expressed as observable instructional components within a short vocational English writing sequence? Second, how can the practice task, independent assessments, and scoring rule be aligned without equating motivation, fluency, and factual task completion? Third, what comparison and analysis would support a bounded educational inference? The contribution is a complete instructional and assessment specification, together with a logical analysis of its scoring and comparison rules. It includes lesson materials, a matched comparison, assessment forms, scoring anchors, implementation records, and explicit interpretations of possible findings. It reports no completed classroom trial and makes no claim of demonstrated effectiveness.

II. Methodological Approach and Design Scope

A. Article Type and Development Procedure

The unit of analysis is the instructional specification: the linked set of goals, activities, materials, assessment responses, scoring rules, and evaluation decisions. Conceptual scholarship requires an explicit account of how theoretical sources contribute to an argument [9]. Here, the sources are used to construct and examine a worked application. Literature selection is purposive rather than systematic; the article does not estimate the frequency or average effectiveness of ARCS interventions. Its conclusions concern the properties and interpretation limits of the supplied design.

Development followed a sequence of specification and internal consistency checks. The target communication task was defined first, including its recipient, factual constraints, and required next action. These requirements informed the practice inquiry, the unfamiliar assessment forms, and the task-fulfillment component of the rubric. ARCS categories were then mapped to teacher prompts, learner choices, and feedback activities. A comparison condition was defined using the same substantive materials, duration, assessment criteria, and feedback categories. The scoring rule and the proposed allocation and analysis were examined for mismatches between the outcome being measured and the claim being made.

The resulting materials were checked for stated content requirements, lesson-time totals, score ranges, form order, and agreement between the prose, tables, and figures. The simple algebraic demonstrations in Section V examine consequences of the declared rules. These checks are inspectable design work. They do not constitute teacher consultation, expert consensus, cognitive interviewing, rater calibration, or a field study. All requests, organizations, prices, and response examples

are fictional instructional materials, and no participant-level observations are reported or synthesized.

The article separates two kinds of conclusion. A design conclusion follows from the written specification, such as the equality of the planned lesson durations or the possibility that different score profiles have the same total. An empirical conclusion would require observations, such as whether learners understand a prompt, whether independent raters apply an anchor consistently, or whether an allocated condition improves performance. This separation permits a substantive methodological contribution while leaving its classroom adequacy open to empirical examination.

B. Relationship to Established Frameworks

Table 1 identifies the function of each principal resource. The comparisons concern its role in the present application; they do not imply that the wider literature lacks the other functions. ARCS informs motivational planning, the CEFR informs the communication objective, constructive alignment connects practice and assessment, and intervention-reporting principles support a reproducible account of delivery [10]. The design contribution lies in specifying how those functions operate together in a bounded writing task, including choices that can create tensions between them.

Design-based research connects the design of learning environments with empirical investigation and refinement [11]. The present contribution is the initial specification and its analytical examination. It can support later classroom iteration without presenting that iteration as already completed. Similarly, Kane's account of validation emphasizes the evidence needed for a stated interpretation or use of scores [12]. In this article, the immediate interpretive target is performance on the specified writing task under documented assessment conditions. Broader uses require additional arguments and observations.

These relationships also reveal a practical tension. Motivational support during learning can be tailored to a learner's needs, while a comparative assessment requires documented and comparable response conditions. The solution adopted here is to permit bounded support during practice and to define common rules for the independent assessments. Another tension concerns informative feedback and treatment differentiation. Both conditions receive feedback and transparent criteria; the comparison varies the complete motivational and interaction package rather than withholding basic assessment support. The resulting estimand is therefore an effect of the specified package, not a pure effect of any single ARCS category.

C. Intended Learners, Setting, and Boundaries

The worked application is intended for vocational learners who can read a short English inquiry, identify explicit information in provider notes, and produce a brief written response. These are instructional prerequisites, not a claim that a particular CEFR level has been established. The materials do not include a validated placement test. Before local use, a teacher should check the prerequisites with a short task separate from the

Table 1: Roles of established frameworks in the instructional specification.

Resource	Function used here	Inspectable application
ARCS [2], [3]	Motivational planning through attention, relevance, confidence, and satisfaction.	Goal elicitation, relevance explanations, supported decisions, and reflection on a revision.
CEFR [1]	Purposeful communication through distinct communicative activities.	A written customer response with a stated recipient, request, constraint, and next action.
Constructive alignment [4]	Coherence between intended performance, learning activity, and assessment.	Common communication functions across practice and unfamiliar assessment forms.
TIDieR [10]	Description of intervention components and their delivery.	Two timed conditions, common materials, teacher instructions, and records of actual exposure.
Design-based research [11]	Theory-informed design examined and refined through empirical work.	An initial specification that can enter classroom development; no completed empirical design cycle is claimed.
Argument-based validation [12]	Examination of the assumptions supporting score interpretations and uses.	Distinct claims about scoring, task-family performance, comparison effects, and intended use.

baseline and outcome forms, documenting any vocabulary or reading barriers that would prevent meaningful participation.

The setting is a teacher-led class with opportunities for individual writing and, in the ARCS-informed condition, paired discussion. Paper materials are sufficient. A digital platform can distribute the same materials, but automated text generation, retrieval, or physiological measurement is not required by the specification. If a later implementation adds such a component, its role, assistance rules, and effect on the comparison must be described as a modification. The educational question should remain interpretable independently of the delivery medium.

The targeted genre is a concise professional email responding to an inquiry whose requested arrangement cannot be met exactly. The learner must explain an alternative without inventing availability, prices, or permissions. Providing the factual notes and final total reduces the need for occupational knowledge and arithmetic unrelated to the writing objective. This constraint improves inspectability, while also limiting authenticity: real workplace exchanges may require clarification, negotiation, policy interpretation, and information gathering across several messages. The current task is a controlled practice opportunity within that broader activity, not a comprehensive measure of vocational competence.

III. ARCS-Informed Instructional Framework

A. Language Outcomes and an Explicit Evidence Pathway

The proposed framework distinguishes the communicative activities relevant to vocational English. Listening and reading contribute to reception; speaking and writing contribute to production. Interaction involves responding to another participant, while mediation includes explaining information or helping others construct meaning. Vocabulary, grammar, pragmatic choices, and contextual knowledge support these activities. This instructional distinction is not a factor model and does not imply that a single assessment measures every aspect of language competence equally well.

For the proposed application, the central learning objective is to produce a clear and appropriate response to an unfamiliar workplace inquiry. Supporting objectives include identifying the request, selecting relevant facts, choosing an appropriate register, organizing the response, and revising language after feedback. Defining the outcome in terms of performance makes

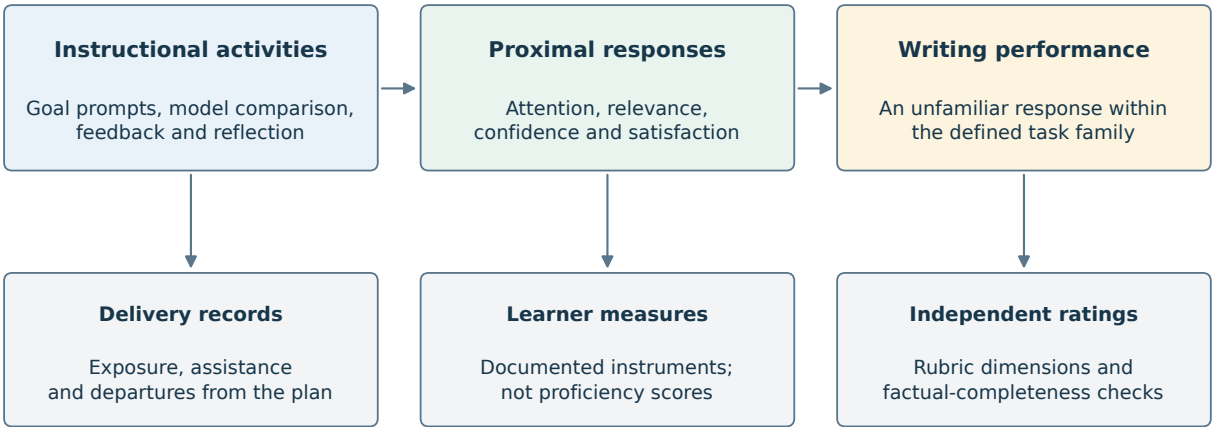
it possible to distinguish an enjoyable lesson from successful transfer to a new task. A student may enjoy a discussion while still misunderstanding the customer's request; conversely, a student may improve a response without reporting a large change in satisfaction.

Figure 1 connects instructional choices, proximal learner responses, and independently assessed performance. The arrows express hypotheses for evaluation, not estimated causal effects. Attention may increase opportunities to engage with a task; relevance may encourage learners to invest effort; confidence may support persistence after an error; and satisfaction may encourage continued participation. Whether these responses improve performance remains an empirical question. Delivery records, learner reports, and independent response ratings provide different observations at the corresponding stages. Their separation allows the proposed relationship to be examined when the observations agree and when they diverge.

B. Motivational Design and Teaching Strategies

The ARCS categories are complementary planning considerations. A teacher can revisit task relevance while adjusting feedback or difficulty, rather than applying the categories as a compulsory sequence. Figure 2 translates these considerations into a lesson sequence with an independent transfer assessment. Tables 2–5 specify local instructional adaptations for attention, relevance, confidence, and satisfaction. Their labels organize this proposal and are not a reproduction of Keller's original subcategory coding or a validated ARCS questionnaire. Each strategy is stated as an instructional action whose delivery could be documented during an evaluation. The tables form a design menu; the selected components of the worked version are fixed by Table 7 and Appendix B. Other entries represent possible documented adaptations, not additional requirements silently added to the two lessons.

Attention strategies emphasize purposeful variation. A short customer message can introduce a communication problem, followed by an individual interpretation, a paired comparison, and a response-writing task. Variation should serve the learning objective rather than compete with it. Humor can make an example memorable when it remains appropriate to the learners and subject matter, but an entertaining anecdote unrelated to the communication problem provides little evidence of



Horizontal arrows are hypotheses, not estimated effects or mediation paths.

Figure 1: Proposed evidence pathway for the writing lesson. Horizontal arrows express hypotheses; the lower row identifies distinct observations. No effects or mediated relationships have been estimated.

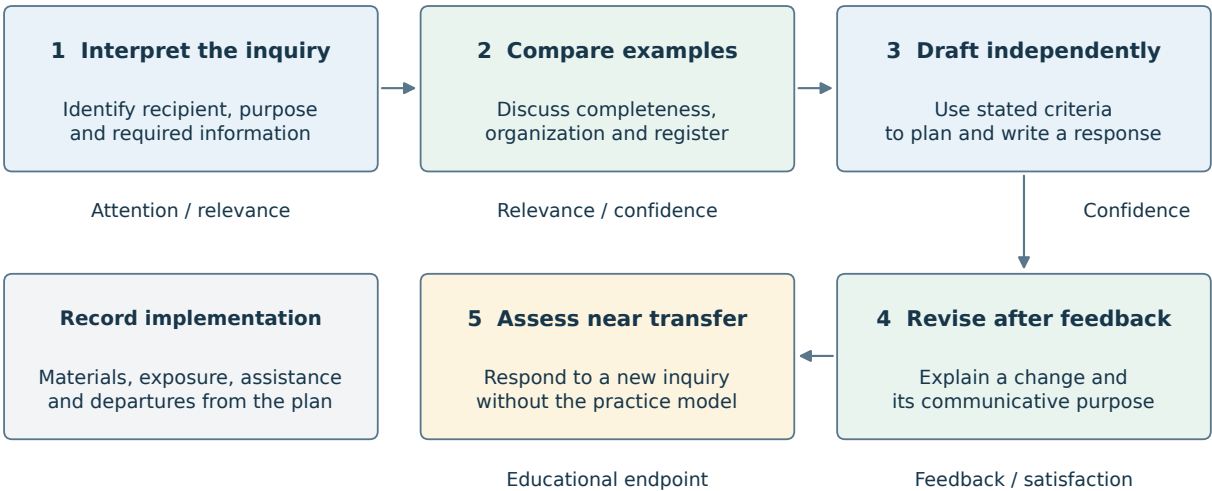


Figure 2: Proposed lesson sequence and evidence collection. Supported practice is separated from an unfamiliar transfer task. The annotations identify intended motivational functions; they are not measured effects.

instructional relevance. The proposed implementation records the activity, its purpose, and the learners’ opportunity to respond.

Relevance strategies in Table 3 connect the task with learners’ experience and anticipated communication demands. The teacher can ask students to identify a workplace or everyday situation in which a similar response would be useful. Students should be able to explain why their selected information matters to the intended recipient. This explanation makes relevance observable as an articulated connection, while avoiding the assumption that every learner has the same occupational plans. Alternative examples can address different vocational interests while retaining comparable language demands and assessment

criteria.

Confidence strategies in Table 4 combine transparent criteria with manageable challenge. Before drafting, learners examine what a successful response needs to accomplish, including the requested information and appropriate tone. Support can include a planning outline or a worked example, followed by a new task on which the learner makes independent choices. Feedback should identify a correctable aspect of performance and allow another attempt. Self-reported confidence and assessed competence are recorded separately because a learner can feel confident while making errors, or produce a competent response while remaining uncertain.

Satisfaction strategies emphasize meaningful progress and

Table 2: Attention: contextual teaching strategies adapted from the ARCS framework.

Local component	Instructional action	Observable implementation evidence
A1. Presentation	Vary vocal emphasis, pauses, examples, and visual presentation to clarify the communication problem.	Lesson materials identify the purpose of each change in presentation.
A2. Activity variation	Alternate individual interpretation, teacher questioning, paired discussion, and independent drafting.	Observation records show which learners have an opportunity to respond.
A3. Media and layout	Use appropriate messages, forms, audiovisual examples, white space, and clear typographic hierarchy.	Each medium is linked to a defined language objective; decorative changes are not scored as learning.
A4. Pacing	Adjust task pacing and assistance in response to observed difficulty while preserving a comparable learning objective.	The teacher records changes in time, task complexity, and support.
A5. Humor	Use respectful, relevant humorous examples or analogies when they clarify a language choice.	Learners can explain the connection between the example and the communication task.

Table 3: Relevance: contextual teaching strategies adapted from the ARCS framework.

Local component	Instructional action	Observable implementation evidence
R1. Experience	Connect the inquiry or response task with existing skills, familiar situations, and suitable analogies.	Learners identify a relevant prior experience or explain a task connection.
R2. Future use	Relate the communication purpose to possible employment, examination, or everyday needs.	Learners explain why particular information matters to an intended recipient.
R3. Learner goals	Offer contexts reflecting different vocational interests while retaining comparable language demands.	Task versions and their common performance criteria are documented.
R4. Cooperation	Provide supportive opportunities for discussion and alternative ways of planning a response.	Participation, assistance, and the selected planning approach are recorded without assuming an effect.

Table 4: Confidence: contextual teaching strategies adapted from the ARCS framework.

Local component	Instructional action	Observable implementation evidence
C1. Criteria	Explain what successful task fulfillment, organization, register, and language control require.	Learners have the rubric before practice and can identify a criterion in an example.
C2. Supported challenge	Use a worked example or planning outline, then provide a new task requiring independent choices.	Records distinguish assisted practice from independent performance.
C3. Attribution	Discuss correctable reasons for success or difficulty, including strategy choice and interpretation.	Learners explain an error and propose a concrete revision.
C4. Progress	Recognize successful task completion while distinguishing perceived confidence from assessed competence.	Confidence responses and performance scores are collected and reported separately.

informative feedback. Hattie and Timperley distinguish feedback about the task and learning process from less informative responses directed only at the person [13]. Recognition is linked to an observable improvement, such as making a request clearer or correcting an inappropriate register. The teacher can invite learners to compare drafts and explain the purpose of a revision. Praise alone is not an outcome measure, and frequent praise should not conceal persistent misunderstanding. The practical intention is to make progress understandable and to preserve opportunities for continued practice. Whether these arrangements sustain motivation must be tested through suitable measures rather than inferred from the strategy descriptions.

C. Activities and a Worked Instructional Specification

Activities can be grouped by instructional purpose: accessing information, examining examples, applying knowledge, and explaining or revising a response. This grouping assigns no fixed learning-retention percentages. The provenance and evidential basis of commonly reproduced learning-pyramid percentages have been questioned in the literature [14]. Here, activity selection is justified by the performance being practiced and the feedback it permits. Listening, reading, discussion,

demonstration, and practice can each be useful when their role in the task is explicit.

A worked lesson specification begins with an unfamiliar inquiry containing a clear request and relevant contextual details. Learners first identify the recipient, purpose, and information required. They then compare two possible responses and discuss how organization and register change the message. Each learner drafts a response independently, receives feedback against the stated criteria, and submits a revised version with a brief explanation of selected changes. Finally, different inquiries assess immediate and delayed near transfer without access to the worked response. These stages are proposed activities; no classroom implementation of this sequence is reported here.

For reproducibility, the lesson materials should include the inquiry, comparison responses, planning prompts, assessment rubric, and feedback instructions. The teacher records the planned duration, actual exposure, assistance provided, and departures from the lesson specification. A comparison condition should cover the same language content and provide comparable time and access to materials, while differing in the explicitly specified motivational design. Describing the conditions in terms of actual components is more informative

Table 5: Satisfaction: contextual teaching strategies adapted from the ARCS framework.

Local component	Instructional action	Observable implementation evidence
S1. Meaningful use	Provide opportunities to apply a language choice in a purposeful communication task.	The task requires a response relevant to a recipient and context.
S2. Recognition	Acknowledge specific progress or successful application, with attention to individual learners.	Feedback identifies an observable feature of the response rather than general ability.
S3. Actionable feedback	Give timely information about the task and a practicable next step for revision.	Feedback and the resulting revision can be compared.
S4. Support schedule	Provide more assistance during initial practice and adjust its timing as learners become independent.	The amount and timing of help are recorded; reduced support is distinguished from reduced teaching time.
S5. Reflection	Invite learners to explain improvements and remaining difficulties without requiring uniformly positive responses.	Draft comparisons and learner explanations document both progress and unresolved problems.

than labeling one class “traditional” and the other “innovative.” The distinction permits a later evaluation to identify what was delivered and what difference the comparison addresses.

D. Learner Goals and Local Adaptation

Examination preparation, employment-related communication, and everyday communication can motivate different learners within the same class. The orientation prompt allows a learner to identify a relevant use without requiring a particular career plan. These preferences inform the explanation of the task; they do not change the facts supplied in the common assessment or the standard used to score a response. A learner who does not identify with the occupational example still receives a clear communication purpose and the same opportunity to practice.

Prior experience can also change task difficulty. Familiarity with booking systems, customer-service conventions, or formal email language should not be treated as a measured instructional effect. The local prerequisite check should identify barriers to reading the notes or understanding the response format before the study begins. Adaptations are documented at the level of the actual prompt, assistance, and task demand, so that subsequent readers can judge what population and implementation the findings address.

IV. Prospective Evaluation Protocol

A. Outcomes and Assessment

The proposed evaluation targets the effect of offering a specified ARCS-informed lesson sequence on performance in an unfamiliar business English writing task. The proposed primary endpoint is an independently scored response 24–48 hours after the second lesson, adjusted for baseline performance and assessment form. A delayed near-transfer task is a secondary endpoint intended to examine whether performance extends beyond the immediately practiced examples. These endpoints are proposals for a future study and should be finalized before recruitment; motivational responses and attendance are not substitutes for them.

Table 6 connects the proposed constructs with tasks, observations, and interpretation limits. The writing rubric separately addresses task fulfillment, organization, register, and language control. Scoring anchors and calibration examples should be finalized before raters see trial outcomes. In the worked specification, two trained raters, masked to condition and

assessment occasion where feasible, independently score every available assessment response. Agreement and the handling of disagreements should be reported alongside the outcome analysis. For a suitable continuous rubric total, an intraclass correlation coefficient should specify the rater model, absolute agreement or consistency, and single-rating or average-rating unit, with its uncertainty interval [15]. Rater agreement does not by itself establish that the rubric measures the intended communicative outcome. The worked rubric in Appendix A states an equal-weight rule, whose suitability for the intended outcome still requires examination.

Motivational outcomes should be assessed with an appropriate instrument whose language, content, response options, and scoring are documented. Adaptation to the local educational setting should include checking whether students understand the questions as intended. Measures of attention, relevance, confidence, and satisfaction should not automatically be forced into a single total score. Alamer and colleagues discuss the methodological distinction involved in confirmatory composite analysis in second-language research [16]; whether such a model is appropriate depends on the construct definition and item design. No motivational questionnaire, composite measurement model, or latent structure is validated in this article.

Activity completion and attendance provide process information, while written responses provide performance information. They should remain separate in reporting even when their associations are explored. Interviews or short learner explanations can investigate why an activity was useful or difficult, but they do not replace a comparison of outcomes. A convincing evaluation can therefore describe successful delivery, motivational response, and performance as distinct findings, including cases where these findings diverge. This arrangement makes an unfavorable or mixed result interpretable instead of forcing every measure into a single claim of success.

B. Intervention Delivery and Comparison

A future implementation should specify the lesson sequence, materials, instructor preparation, delivery mode, duration, tailoring rules, and expected learner activities. The TIDieR reporting guide provides a useful structure for documenting intervention components and their actual delivery [10]. Applying those principles here would mean supplying the

Table 6: Prospective mapping of constructs, measurements, and interpretation limits.

Construct	Proposed measurement	Timing or unit	Interpretation boundary
Writing performance	Unfamiliar inquiry; rubric for task fulfillment, organization, register, and language control.	Comparable baseline and end-of-sequence tasks; independently rated student response.	Primary educational endpoint; requires scoring reliability and appropriate class-level comparison.
Transfer	A different inquiry without access to the practice response.	Prespecified delayed assessment; student response.	Assesses application beyond immediate examples; does not establish general occupational competence.
Motivational response	Appropriate, locally checked measures of attention, relevance, confidence, and satisfaction.	Prespecified occasions; student responses.	A motivational response is distinct from language proficiency.
Engagement and delivery	Attendance, opportunities to respond, completed drafts, assistance, and feedback records.	Lesson and class-level process records.	Describes exposure and implementation; does not establish effectiveness.

communication tasks and feedback rubric, recording assistance, and distinguishing the planned lesson from what occurred. Such documentation should cover both the ARCS-informed and comparison conditions.

Fidelity assessment should focus on observable components: whether the teacher explained task relevance, supplied success criteria, provided actionable feedback, and allowed an independent transfer attempt. A short observation form can record these components without treating their presence as proof of effectiveness. Teacher reflection and classroom observations can also identify unintended differences, including additional time, resources, or feedback in one condition. These differences matter because an apparent benefit could arise from greater exposure rather than the motivational organization of otherwise comparable instruction.

C. Worked Delivery Specification and Testable Contrast

The worked specification in Table 7 uses two 45-minute lessons, followed by a 20-minute writing assessment 24–48 hours after the second lesson. A baseline task precedes the first lesson, and a delayed near-transfer task is proposed 14 days after the second lesson, with a prespecified scheduling window of three days on either side. These timings are design choices for an illustrative implementation, not an observed exposure, an established optimal dose, or a registered trial commitment. A feasibility study should examine whether the tasks and timing are appropriate before an effectiveness comparison is finalized.

Appendix A supplies the practice inquiry, model responses, assessment forms, scoring anchors, and record fields. The comparison condition is described explicitly because a label such as ordinary instruction does not define what learners receive. Both conditions retain transparent criteria and access to feedback; withholding those essentials would change the learning opportunity in a way unrelated to the intended comparison. Structured goal elicitation, relevance explanations, planned reflection, and paired consideration of alternatives differentiate the proposed package. Some motivational support may also occur in the comparison condition, and observation should document that overlap.

The primary contrast concerns assignment to this entire instructional package versus assignment to the specified comparison. It is not the effect of ARCS as an abstract theory or of any isolated activity. Teacher contact, peer interaction, and

perceived novelty could contribute to a package difference. Component-specific claims would require a further design that varies those components independently. The protocol therefore records their delivery while preserving a modest interpretation of the main comparison.

Baseline and immediate follow-up assessments use Forms A and B in counterbalanced order. Within each participating class, students are assigned to A-then-B or B-then-A sequences by a prespecified balanced allocation, independently of the class treatment assignment. A student receives each form once. This arrangement prevents a fixed form difference from being identical to assessment occasion. It does not establish that the forms are psychometrically equivalent, so their difficulty, completion time, and interpretation still require piloting. Form C is reserved for delayed transfer and is analyzed as a separate endpoint rather than assumed to share an identical scale difficulty.

Two raters independently score every available assessment response using the four dimensions in Table 10. The proposed primary score is the mean of their total scores, each ranging from 0 to 16. A response with only one rating awaits the second independent rating; if that rating cannot be obtained, its primary composite remains missing rather than changing to a single-rater endpoint. Singly rated responses and the reason for incomplete rating are reported. Scripts are assigned anonymous identifiers, presented in a shuffled order, and stripped of class, condition, learner name, and occasion information where possible. Raters necessarily see the inquiry form and any language that could reveal a practiced strategy; complete masking cannot be guaranteed. Disagreements are retained for reliability analysis. A consensus discussion may support training, but should not replace the independent ratings used to estimate agreement.

The expected educational claim is correspondingly specific: the proposed package may improve independently rated performance on an unfamiliar writing task. Its evaluation would report the estimated condition difference and uncertainty, together with a prespecified educationally meaningful threshold. A motivational increase without a corresponding performance benefit would support a different conclusion from improvement in both. A performance benefit would not, by itself, establish that a particular ARCS component mediated that benefit. The assessment and delivery records make these alternative findings

Table 7: Worked two-lesson specification. Each condition receives 45 minutes per lesson.

Session and activity	Minutes	ARCS-informed condition	Content-matched comparison condition
Lesson 1: orientation	5	Elicit a learner goal and connect the task with a possible communication need.	State the topic and lesson objectives; learners read the same task instructions.
Inquiry analysis	10	Ask learners to identify the recipient's priorities and explain their relevance.	Explain the inquiry's content; learners answer the same factual comprehension questions.
Model comparison	10	Compare the two supplied responses and justify choices using the rubric.	Explain the same models and rubric; learners identify the specified features individually.
Independent draft	15	Use the common task; invite a brief plan against the stated success criteria.	Use the common task and criteria for individual drafting.
Closing activity	5	Record one successful choice and one manageable next step.	Complete a factual review of the language points covered.
Lesson 2: recap	5	Revisit the learner goal and the planned improvement.	Review the lesson topic and previously explained language points.
Feedback and revision	15	Apply the common feedback and explain the communicative purpose of a change.	Apply the same feedback categories to correct the draft.
Consolidation	10	Discuss alternative formulations with a partner, referring to the recipient and criteria.	Complete individual exercises on the same formulations and language points.
Independent redraft	10	Submit a new version of the practice response.	Submit a new version of the same practice response.
Closing activity	5	Explain specific progress and a remaining difficulty.	Summarize the language points and completed corrections.

Both conditions receive the same practice inquiry, factual notes, model responses, rubric, total lesson time, and feedback categories. The comparison estimates the effect of the complete specified package, including its interaction pattern; it does not isolate each ARCS category.

interpretable without requiring every observed response to fit the same explanation.

Before comparative recruitment, a feasibility stage should check whether the instructions are understood, the work can be completed in the allotted time, the scoring anchors can be applied consistently, and the two conditions are distinguishable in practice. The decision criteria and actions should be written before that stage begins. Persistent misunderstanding calls for prompt revision; disagreement about rubric anchors calls for rater calibration; indistinguishable delivery calls for a clearer intervention contrast. The proposal supplies these decision categories, but does not assert that any feasibility criterion has already been satisfied.

D. Allocation, Precision, and Statistical Analysis

When instruction is delivered to intact classes, a prospective comparison should include multiple independently allocated classes per condition. Allocation should occur after the eligible classes and baseline assessment procedures are defined. The design should address teacher effects, shared materials, and possible contamination between conditions, with analysis and reporting aligned to cluster allocation [17]. If teachers are confined to a single condition, the resulting association between teacher and condition must be addressed through design rather than ignored during analysis. Recruitment, allocation, losses, and analyzed observations should be reported at both class and student levels. General trial reporting can follow CONSORT 2025, supplemented by design-specific cluster guidance, with explicit reporting of registration, protocol access, and deviations [18]. These reporting frameworks are used as methodological resources; they do not turn the present proposal into a completed trial.

The required number of classes depends on the smallest educationally meaningful difference, outcome variability, intra-class correlation, class-size variation, anticipated missingness,

and desired precision. No empirical planning estimates are supplied by the present design work, so a numerical sample size is not inferred from its worked examples. Planning should use a defensible external estimate or feasibility information and examine sensitivity to plausible assumptions. Adding students to two classes is not equivalent to adding independent classes. The planning document should state the chosen assumptions and their sources before the effectiveness evaluation begins.

For an appropriately designed study, one possible baseline-adjusted model is

$$Y_{ij,1} = \beta_0 + \beta_1 T_j + \beta_2 Y_{ij,0} + \beta_3 F_{ij,1} + u_j + \varepsilon_{ij}, \quad (1)$$

where i identifies a student, j a class, T_j the allocated condition, $F_{ij,1}$ the follow-up form indicator, and u_j a class effect. The outcome is the mean of the two independent rater totals. A form term is included because counterbalancing does not establish equal difficulty. Eq. (1) is a prospective specification, not a model fitted in this article. Its use requires an adequate number of clusters, a suitable outcome scale, and an analysis consistent with the actual allocation and teacher structure. Small-sample corrections or an alternative class-level analysis should be chosen when the design warrants them. The model alone cannot repair confounding or insufficient independent replication. The learner-average estimand in Eq. (7) must also govern estimator selection. Participant-average and class-average effects can differ when class size is related to outcomes or effect size, and a mixed-effects coefficient need not estimate either target without appropriate assumptions [19]. For the stated learner-average target, the protocol should prespecify an independence estimating-equation analysis with equal learner weights, baseline and form adjustment, and class-robust uncertainty with a suitable small-sample correction. The mixed-effects formulation is an alternative or sensitivity specification only when its assumptions and target are justified. The number of independent classes remains a limitation for either approach.

The primary analysis should follow the prespecified allocation-based estimand and report an effect estimate with an uncertainty interval in interpretable score units. Secondary outcomes should be identified as such, with a stated approach to multiplicity. Missing outcomes require reporting their amount and reasons, followed by a justified analysis and sensitivity assessment; complete cases should not silently become the target population. Subgroup and mediation analyses should be undertaken only when their assumptions and precision are supportable. An equivalence claim would require prespecified, substantively justified bounds and an appropriate equivalence analysis [20]. A nonsignificant result should not be recast as proof of equivalence, and an improvement in an attitude score should not be substituted for the primary performance endpoint.

E. Implementation, Records, and Ethics

Appendix B provides the administration guide, common teacher materials, and prompts distinguishing the conditions. Figure 3 displays the planned assessment order. Baseline precedes class allocation; the within-class form sequence is assigned through a separate balanced procedure. Assessment exposure and actual timing are recorded even when a learner misses a lesson or an assessment. Keeping the original assignments permits missingness and deviations to be examined without redefining the comparison after outcomes are known.

A future research implementation requires the appropriate institutional ethics determination and documented consent arrangements before recruitment and data collection. Ordinary teaching participation should be distinguished from consent to use responses for research, with alternatives or withdrawal arrangements specified as appropriate. Collect only the identifying information needed to administer and link the observations, and store the identity key separately from the scoring dataset. Raters should receive anonymous scripts and the task information needed to interpret them. These procedures describe the prospective study; no approval or consent for an unperformed study is claimed.

V. Analytical Examination of the Specification

A. Alignment of Practice and Assessment Content

Table 8 maps the practice and assessment forms to their shared communication requirements. Each assessment asks the learner to handle a requested arrangement, a constraint, a supplied service, a total price and deadline, and a decision about an alternative. The practice form differs in surface detail and in the kind of constraint: it concerns completing an order in time rather than accepting a revised booking time. These shared functions make the intended near-transfer relationship explicit. They do not establish equal difficulty or statistical interchangeability between the forms.

The mapping supports a limited design conclusion: every form supplies the information needed to construct a response addressing the intended functions. It also makes possible differences visible. The service details vary in number and familiarity, the practice example permits a delivery-address request, and the transport task does not require changing both

departure and return. These differences may affect cognitive demand, vocabulary use, or completion time. Piloting must examine them rather than assuming that similar formatting produces equivalent measures.

The assessment forms should remain unavailable during practice. Learners can receive feedback on the common communication functions without seeing the particular booking details used at follow-up. This preserves the distinction between learning a response procedure and rehearsing the exact assessed message. The design does not attempt to eliminate all learning from baseline assessment. Instead, both conditions receive the same baseline procedure, and the resulting comparison concerns the instructional packages in the presence of that common assessment experience.

B. Score Aggregation and Functional Completeness

Let x_{ird} be the score assigned to response i by rater r on dimension d , where $r \in \{1, 2\}$ and the four dimensions are task fulfillment, organization, register, and language control. Each dimension receives an integer from 0 to 4. The specified total and the primary score are

$$S_{ir} = \sum_{d=1}^4 x_{ird}, \quad Y_i = \frac{S_{i1} + S_{i2}}{2}, \quad 0 \leq Y_i \leq 16. \quad (2)$$

Eq. (2) defines a transparent composite with equal numerical weights. It does not establish that the distances between rubric categories are psychometrically equal or that the four dimensions are interchangeable. A future analysis should examine whether treating the total as approximately continuous is defensible for its purpose, with dimension-level and appropriately specified sensitivity analyses available when the observed distribution warrants them.

A simple counterexample shows why the total cannot be read as a direct measure of complete communication. Consider the constructed single-rater profiles (1, 4, 4, 4) and (4, 3, 3, 3). Both sum to 13, although the first profile has substantially weaker task fulfillment. These vectors are arithmetic examples, not scores assigned to learners or validated benchmark scripts. They establish a property of the aggregation rule: strength in three dimensions can compensate numerically for weakness in the fourth. They do not describe how often such profiles occur, which profile teachers would prefer, or whether independent raters would assign those exact values to any response.

The relevant bound follows immediately. If a single-rater total is S and its task-fulfillment score is x_T , the remaining three dimensions contribute at most 12 and at least zero. Therefore,

$$\max(0, S - 12) \leq x_T \leq \min(4, S). \quad (3)$$

For example, a total of 12 is mathematically compatible with a task-fulfillment score anywhere from 0 to 4. Eq. (3) is an elementary scoring implication, not a new psychometric theorem. The supplied verification script exhaustively checks the bound over the $5^4 = 625$ permitted single-rater profiles. Enumeration checks the declared arithmetic only; those profiles

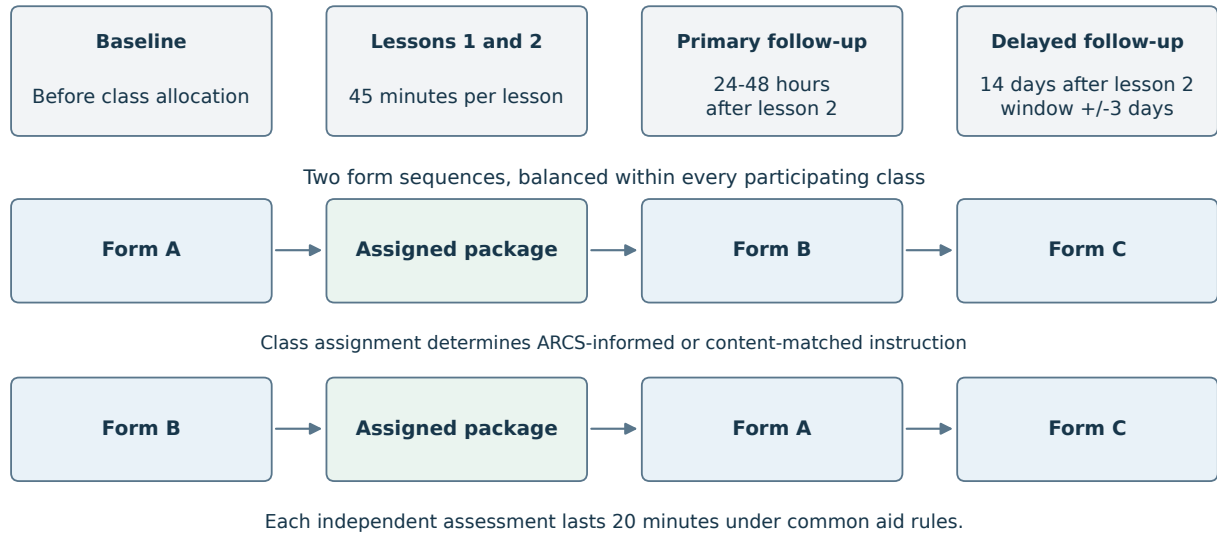


Figure 3: Prospective allocation and assessment order. Within each class, learners receive either A then B or B then A; both sequences receive delayed Form C. The timeline represents planned procedures, not completed recruitment or data collection.

Table 8: Content alignment across the fictional practice and assessment forms.

Communication requirement	Practice P	Form A	Form B	Form C
Acknowledge request	24 welcome packs for Friday.	Room for 12 on Monday.	Eight laptops on Tuesday.	Transport for ten on Wednesday.
Explain constraint	Four packs ready Wednesday; complete delivery depends on timely confirmation.	09:30–12:30 replaces 09:00–12:00.	13:30–16:30 replaces 13:00–16:00.	08:30 departure replaces 08:00; return remains 17:00.
Confirm supplied service	Notebook and pen in every pack.	Projector; refreshments at 10:30.	Chargers; delivery before 13:30.	Ten seats and luggage space.
State total and deadline	296; Tuesday 16:00.	96; preceding Thursday 15:00.	104; preceding Friday 12:00.	150; Monday 10:00.
Request next decision	Confirm the complete order.	Accept alternative room time.	Accept alternative equipment time.	Accept alternative departure.

are not equally likely student outcomes and are not a simulated classroom sample.

To retain the factual component of the communication objective, the protocol records five content checks alongside the rubric. For rater r , let $c_{irk} = 1$ indicate that response i accurately fulfills required function k , and let $p_{ir} = 1$ indicate a material unsupported promise. A descriptive functional-completeness indicator is

$$Q_{ir} = (1 - p_{ir}) \prod_{k=1}^5 c_{irk}. \tag{4}$$

Thus, $Q_{ir} = 1$ requires all five functions and no material unsupported promise. Eq. (4) is a content rule for this task family. It is not a validated proficiency cutoff and does not replace the primary composite. Raters record the checks independently; reporting should identify disagreements rather than silently converting them into consensus. A cautious descriptive summary can report the proportion of assessable scripts meeting the rule according to both raters, together with the number for which their classifications differ.

A material unsupported promise is a commitment that conflicts with or goes beyond the provider notes in a way that changes what the customer can rely on. Promising the requested 09:00 room start when only 09:30 is available is material. Asking the customer to confirm an address is a request, not an unsupported promise. A spelling error that leaves a fact

unambiguous concerns language control rather than factual correctness. These distinctions help prevent a single error from indiscriminately determining every dimension while preserving the importance of accurate commitments.

The composite and the content checks answer complementary questions. A higher composite indicates better performance under the declared weighting rule; the content checks show whether the practical communication requirements were fulfilled. An analysis should not change the primary endpoint after seeing which summary favors a condition. Nor should it describe all improvement as success if inaccurate commitments become more common. Reporting the dimensions and factual checks permits that tension to be examined directly.

C. What Counterbalancing Resolves

Forms A and B are assigned in opposite baseline/follow-up sequences within every class. To illustrate the purpose of this arrangement, suppose a score has an additive decomposition into a learner component a_i , an occasion component g_t , and a form component h_f . Under that illustrative model, the expected changes for the two sequences are

$$\begin{aligned} E(\Delta_{AB}) &= (g_1 - g_0) + (h_B - h_A), \\ E(\Delta_{BA}) &= (g_1 - g_0) + (h_A - h_B). \end{aligned} \tag{5}$$

The equally weighted mean of these expected changes cancels a fixed additive form difference:

$$\frac{E(\Delta_{AB}) + E(\Delta_{BA})}{2} = g_1 - g_0. \quad (6)$$

Eqs. (5) and (6) explain a design property under the stated assumptions. They neither estimate instructional improvement nor prove that the forms are equivalent. Differential practice, learner-by-form interactions, condition-specific form effects, or unequal loss across sequences can break the simple cancellation argument. The proposed outcome model therefore retains a form term, and sequence counts and missingness should be reported in each condition.

When class size is odd, the balanced allocation can differ by one learner, with the extra sequence assignment determined by the prespecified allocation procedure. Equal allocation refers to the plan, not a guarantee of equal numbers with observed outcomes. Analysis should preserve the assigned sequence information and account for the actual design. A fixed-effect term cannot repair an assessment form that measures a different construct or is unintelligible to a substantial subgroup; those problems require revising the assessment before the comparative study.

Form C changes the practical context and assessment delay together. Its outcome can address whether the condition difference is also present on that delayed task. A raw difference between the immediate and delayed means cannot isolate forgetting, because task form and elapsed time change simultaneously. The protocol therefore treats the delayed outcome as a separate secondary endpoint. A study specifically targeting retention decay would need additional counterbalancing or linked tasks designed for that purpose.

D. The Instructional Contrast and Its Interpretation

The proposed primary target is the average effect of assignment to the complete ARCS-informed package for the learners enrolled before class allocation. If there are N such learners across the participating classes and $Y_{ij}(z)$ is the specified follow-up outcome under class assignment z , the finite-population target is

$$\tau = \frac{1}{N} \sum_j \sum_i \{Y_{ij}(1) - Y_{ij}(0)\}. \quad (7)$$

Eq. (7) defines the target; it is not an observed result. Its interpretation presumes adequately specified condition versions and a defensible account of interference between classes. Peer interaction within a class is part of the package. Sharing materials or instructional prompts between allocated classes can change the contrast and should be examined in implementation records. The statistical analysis must address class allocation, teacher structure, and missing outcomes to support an estimate of this target.

The intervention varies goal elicitation, explanations of task relevance, planned next steps, reflection, and paired consideration of alternative formulations. Consequently, an observed package difference cannot identify the separate

contribution of attention, relevance, confidence, satisfaction, or peer interaction. Measuring motivation after instruction would not by itself resolve that identification problem. A claim about a mechanism would require a suitable additional design and assumptions, with adequate measurement and temporal ordering. The present package is deliberately evaluated as a whole.

Table 9 connects possible observations with their bounded interpretations. The rows are hypothetical patterns, not expected results or an empirical results table. They are useful because a methodological specification should remain interpretable when its outcomes disagree. A successful lesson delivery may yield no performance benefit; an apparent benefit may depend on a particular form; an improved composite may coexist with persistent factual errors. Each pattern raises a different question and calls for a different next analysis or design decision.

The analytical examination yields four directly inspectable conclusions: the materials specify common communication functions, the total score permits compensation between dimensions, counterbalancing removes a fixed form contrast only under stated assumptions, and the primary comparison targets a complete instructional package. These conclusions establish the internal logic and boundaries of the design. Evidence about usability, score interpretation in a population, and comparative effectiveness remains the purpose of subsequent empirical work.

VI. Discussion

A. Methodological Contribution

The article's contribution is the joint specification of an instructional package, its comparison, and the meaning of its assessment outcomes. The familiar ARCS categories become observable prompts and learner activities; constructive alignment becomes a content mapping across practice and unfamiliar tasks; and the assessment argument becomes a set of explicit score rules and interpretation limits. This integration produces materials that can be inspected, adapted, and tested. It is an applied methodological contribution, not a new general theory of motivation or an estimate of an instructional effect.

The analytical results sharpen the practical design. The score example shows that a favorable composite can coexist with weak task fulfillment, so dimension scores and factual checks are retained in the evaluation. The counterbalancing argument explains why assessment form must be addressed, while also identifying what an additive cancellation does not establish. The definition of the package contrast prevents a future effect from being assigned to an isolated component that the design never varies independently. These are consequences of the written specification, with relevance to the interpretation of later observations.

A central practical choice is to preserve transparent criteria and feedback in both conditions. Such support is integral to a meaningful learning opportunity and need not be removed to create a comparison. What differs is the complete organization of goal elicitation, relevance explanation, planned reflection, and interaction. That choice makes the contrast more specific

Table 9: Interpretation of hypothetical evaluation patterns; no outcomes are reported.

Possible observation	Interpretation supported by suitable data	Inference that would remain unsupported
Planned components delivered; writing difference imprecise.	The specified delivery was feasible in the observed setting; the performance comparison remains uncertain.	Delivery fidelity proves effectiveness, or an imprecise null establishes equivalence.
Motivational response improves; writing does not show a clear benefit.	The measured motivational response and writing outcome diverge.	Motivation can substitute for language performance.
Composite improves; factual commitments remain inaccurate.	Performance improves under the specified weighting, with an unresolved functional weakness.	Every aspect of customer communication improves.
Writing difference favors the package after an appropriate class-level comparison.	Assignment to the complete package affects the stated endpoint under the study conditions.	A particular ARCS category or peer discussion alone caused the effect.
Delayed Form C also favors the package.	The difference extends to this delayed task within the defined family.	General occupational competence or an isolated retention-decay rate has been established.

than a comparison with an undefined ordinary lesson. It also makes the causal claim narrower: the design can address the additional package as implemented, while the comparison may contain motivationally supportive elements of its own.

Another choice concerns the relationship between practice and independent assessment. Supported drafting allows learners to discuss alternatives and act on feedback. The follow-up task requires an individual response to different facts without the worked response. This distinction reduces direct copying as an explanation for successful completion, but does not remove every effect of practice or task familiarity. The endpoint is explicitly bounded to a related writing task, and the delayed form adds a later observation within that same task family.

B. Use and Adaptation

The specification can serve two purposes. A teacher can use the materials for formative practice while documenting local adaptations. A researcher can use the defined conditions and assessment plan as the starting point for feasibility work and a subsequent comparative study. These uses require different decisions: ordinary classroom feedback need not create a research dataset, while an effectiveness study requires prospectively documented recruitment, allocation, measurement, and analysis. The materials should not become a high-stakes proficiency test simply because they include a numerical rubric.

Adaptation should preserve the intended communication functions while examining whether a changed context introduces new demands. Replacing a room booking with a technical maintenance inquiry may require specialist knowledge that the current task intentionally avoids. Translating administration instructions can improve access while changing how learners interpret a prompt. Providing a glossary during practice may be compatible with the lesson goal, whereas supplying sentence-level help during the independent assessment changes the response conditions. Documentation allows these decisions to be evaluated rather than hidden under a common intervention name.

A future study should also consider the consequence of the assessment incentives. If learners discover that polished

language can offset an inaccurate promise in the total score, they may optimize the wrong behavior for customer communication. Reporting the factual checks alongside the composite makes that possibility visible. It does not solve every weighting problem, but it allows readers to determine whether a claimed benefit reflects functional improvement, stylistic improvement, or a mixture. Any change to the primary scoring rule should precede the comparative outcome analysis.

C. Limitations and Research Progression

The design has not been implemented with teachers or learners. The proposed rubric has no reported reliability or validity evidence, and the forms have not been calibrated. The literature selection is purposive, and the article does not establish that its particular combination is superior to alternative ARCS implementations. Its restricted genre and short sequence also limit the range of learning claims it can address. These are boundaries of an initial methodological specification, and they should guide its empirical use.

The next empirical stage should examine learner comprehension, task timing, teacher delivery, rater interpretation, and differentiation between the conditions. Appendix C supplies a staged examination plan and the decisions each source of evidence would inform. A study that finds unclear prompts or persistent disagreement should revise the materials before fixing an effectiveness protocol. A feasible implementation can then support a class-level comparison with a prospectively defined estimand, sufficient independent replication, and suitable missing-data and sensitivity analyses.

The completed design work and the outstanding empirical questions should remain separate in reporting. Equality of planned duration, completeness of supplied facts, and correctness of the score arithmetic are directly checkable now. Actual exposure, accessibility, score quality, and educational effects must be evaluated in the intended setting. This distinction permits the article to contribute an inspectable research design while preserving the possibility that later evidence will require substantial revision of its activities or interpretation.

VII. Conclusion

This article develops an ARCS-informed specification for vocational English writing that connects lesson delivery, a defined comparison, independent assessment, and bounded interpretation. It supplies two timed lessons, fictional practice and assessment materials, analytic scoring anchors, factual-completeness checks, administration procedures, and implementation records. Its analytical examination identifies the compensatory nature of a rubric total, the assumptions behind form counterbalancing, and the scope of a package-level instructional contrast.

The contribution is a reproducible design and a clear basis for empirical examination. It does not establish instructional effectiveness, validated proficiency measurement, or broad occupational transfer. By stating what is taught, what differs between conditions, what is scored, and what each result could support, the specification enables a subsequent study to address a coherent educational question. Its value will ultimately be assessed through its use, critique, adaptation, and empirical evaluation in vocational language classrooms.

APPENDIX A. Worked Teaching and Assessment Materials

The materials in this appendix were developed for the proposed design. All organizations, requests, prices, and responses below are fictional instructional examples. They are not participant observations, validated test items, or evidence of an intervention effect. They provide a complete inspectable starting specification that can be piloted and adapted with changes documented.

A. Common Instructions and Practice Inquiry

For each writing task, act as an assistant replying to a customer. Write an email of approximately 100–140 words using only the supplied facts. Identify the request, state what can be provided, explain any constraint, give the relevant cost or timing information, and ask for the customer's next decision. A small departure from the word range is not scored separately; the rubric evaluates communicative performance. Assessment time is 20 minutes. Both conditions use the same access rules for dictionaries or other aids, specified in advance; the worked default is no external assistance. Practice may use models and feedback, but the independent assessments do not.

Practice Form P: welcome packs. A customer writes: "We need 24 welcome packs for a staff event on Friday. Can all packs arrive before noon? Each pack should include a notebook and a pen. Please confirm the total cost and what you need from us before proceeding." The supplier's notes state: each complete pack costs 12 currency units; delivery for the entire order costs 8; 20 packs are ready now; four additional packs will be ready on Wednesday; the complete order can arrive on Friday before noon if the customer confirms by Tuesday at 16:00. The total is 296 currency units. The task requires no independent arithmetic; the supplied total prevents calculation skill from becoming an unintended prerequisite.

Model response P1. "Dear Customer, Thank you for asking about the welcome packs for your Friday staff event. We

can provide 24 packs, each containing a notebook and a pen. Twenty packs are ready now, and the remaining four will be ready on Wednesday. We can deliver the complete order before noon on Friday if you confirm by Tuesday at 16:00. The packs cost 288 currency units in total, and delivery costs 8, giving a final total of 296. Please confirm the order by the stated deadline and send the delivery address so that we can arrange the shipment. Kind regards, Sales Assistant."

Model response P2. "Hi, We have 20 packs now, so everything can go today. The price is 288 with delivery. Let us know if you want them. Thanks." This deliberately flawed example contradicts the available quantity and total cost, omits the confirmation deadline, and does not answer the complete delivery request. It is provided for discussion of task fulfillment, clarity, and register. It is not an actual learner response and is not evidence that instruction caused any particular error.

The teacher gives both conditions the same two responses and the same factual notes. In the ARCS-informed condition, learners explain why a difference matters to the recipient and connect a revision to a stated goal. In the comparison condition, the teacher explains the differences and learners complete the specified identification exercise. Feedback in both conditions addresses an omitted requirement, an inaccurate fact, or a language choice that impedes the intended message. Corrections should not add facts unavailable to the learner.

B. Assessment Forms

Form A: meeting-room booking. A customer asks to book a meeting room for 12 people on Monday from 09:00 to 12:00, with a projector and refreshments, and requests the total price and confirmation deadline. Provider notes: Room Cedar seats 12; it is available from 09:30 to 12:30 rather than the requested times; the projector is included; refreshments are available at 10:30; the room and refreshments together cost 96 currency units; the customer must confirm by the preceding Thursday at 15:00. Reply by explaining the alternative time, confirming the facilities and total, and asking whether the alternative is acceptable before the deadline. No further availability or discount may be invented.

Form B: equipment reservation. A customer asks to reserve eight laptops on Tuesday from 13:00 to 16:00, with chargers and delivery, and requests the total price and confirmation deadline. Provider notes: eight laptops are available; the complete set can be provided from 13:30 to 16:30 rather than the requested times; chargers are included; delivery is available before 13:30; the equipment and delivery together cost 104 currency units; the customer must confirm by the preceding Friday at 12:00. Reply by explaining the alternative time, confirming the equipment and total, and asking whether the alternative is acceptable before the deadline. The structure parallels Form A, but equal difficulty has not been established.

Form C: visitor transport. A customer requests transport for ten visitors on Wednesday, departing at 08:00 and returning at 17:00, with space for luggage, and asks for the total price and confirmation deadline. Provider notes: a vehicle seats ten and has luggage space; it can depart at 08:30 rather than

08:00; return at 17:00 is available; the complete service costs 150 currency units; confirmation is required by Monday at 10:00. Explain the departure constraint, confirm the return and capacity, state the total, and ask whether the alternative is acceptable. This form examines delayed near transfer to another context within the same customer-response task family; it is not assumed to be interchangeable with Forms A and B without evidence.

For content scoring, each assessment has five required functions: acknowledge the request; explain the relevant time constraint accurately; confirm the requested facilities, equipment, or capacity; state the supplied total and confirmation deadline; and request a decision about the alternative. These functions provide a content key rather than a script that a learner must copy. Wording and organization may vary. Raters use the actual response and the task notes to distinguish an acceptable paraphrase from an unsupported promise.

C. Proposed Analytic Rubric

Table 10 defines four separately scored dimensions. Each receives an integer from 0 to 4. A rater's total is the sum of the four dimension scores, and the proposed primary score averages the two independent totals. Equal weighting is an explicit design choice rather than an empirically validated property. Dimension scores should also be reported so that an aggregate change does not conceal an important trade-off, such as accurate facts expressed in language that the intended recipient cannot readily understand.

Rater calibration uses example scripts separate from assessment scripts. Raters first score independently, identify disagreements, and discuss how the anchors apply. Revisions to anchors occur before masked outcome scoring and are versioned. If a script omits factual requirements but uses accurate English, its language-control score can exceed its task-fulfillment score; the dimensions should not all be reduced because of one global impression. Conversely, copying correct facts into an incomprehensible message need not receive a high communication score. The process supports consistent application but does not itself prove construct validity.

The supplied flawed practice response can illustrate this distinction, but it is not assigned a validated benchmark score here. Illustrative score vectors in Section V examine the aggregation rule only; they are not empirical rater judgments or calibration statistics. A feasibility study should examine actual rater disagreements, the distribution of dimension scores, and whether learners and teachers interpret the intended task consistently. Floor or ceiling effects, persistent ambiguous anchors, or very different form difficulty would require revision before an effectiveness study.

D. Implementation Record and Adaptation Rules

For each lesson, record a class identifier, teacher identifier, date, planned and actual duration, material version, completed activity blocks, assistance provided, and any departures from the comparison schedule. For the ARCS-informed condition,

also record whether the learner-goal prompt, relevance explanation, planned next step, and revision reflection were delivered. Record analogous prompts if they occur spontaneously in the comparison condition. These are observations of exposure and overlap, not a score proving that the intervention worked.

For each assessment, record an anonymous learner identifier, class, assigned form sequence, form, occasion, actual assessment time relative to the lesson, aid use, submission status, both independent sets of dimension ratings, five content checks per rater, and any material unsupported promise. Keep the identity key separately from the scoring file. Distinguish nonattendance, refusal, lost work, and unusable files where those reasons are known; do not infer a reason from a blank cell. Corrections to data entry should retain an audit record. No example row containing a purported participant observation is supplied.

Adaptation can change the occupational surface context, vocabulary support during practice, or delivery format while preserving the stated communication functions. Changes to assessment assistance, core task constraints, time limits, scoring anchors, or the intervention contrast alter the evaluated specification and must be documented. If an adaptation changes the construct or difficulty substantially, the original rubric and form assumptions require renewed examination. This rule permits contextual use without treating every locally modified lesson as an equivalent implementation.

APPENDIX B. Teacher and Assessment Administration Guide

A. Preparation and Common Classroom Materials

Prepare one copy of Practice Form P, its supplier notes, both model responses, and the rubric for every learner. Provide space for an initial draft, a revised draft, and the short activity response specified by the allocated condition. A teacher needs the factual key and feedback guide before the lesson. Record a version identifier on the teacher packet so that a corrected prompt or changed example can be distinguished from an earlier version. Assessment Forms A, B, and C are stored separately from the teaching packet.

Before delivery, the teacher reads both conditions and identifies their common components and differences. Preparation should cover the factual content, the four scoring dimensions, the planned durations, and the rules on assistance. The purpose is to support consistent implementation, not to persuade the teacher that one condition is expected to succeed. Where teachers deliver more than one class, the research plan should address the resulting teacher structure and the possibility that strategies cross between conditions. A single teacher in one condition should not silently stand in for the entire instructional approach.

Common factual comprehension questions are: How many packs are requested? What must each pack contain? When must the full order arrive? How many packs remain to be completed, and when will they be ready? What total must the customer pay, and when is confirmation needed? The supplied answers are 24; a notebook and a pen; Friday before noon; four, ready Wednesday; and 296 currency units with confirmation

Table 10: Proposed 0–4 scoring anchors for an assessable workplace email. The rubric is not yet validated.

Score	Task fulfillment	Organization	Register	Language control
4	All five required functions are fulfilled accurately, with no material unsupported promise.	The request, explanation, and next action are easy to locate and logically connected.	Tone and forms of address are consistently suitable for a professional customer exchange.	Language communicates the intended meaning clearly; occasional slips do not hinder interpretation.
3	The main request and constraint are handled accurately; one minor omission remains.	The response is coherent, with a minor sequencing or linking weakness.	Tone is generally appropriate, with an occasional awkward or overly informal expression.	Errors occur, but the intended message remains clear without substantial reader effort.
2	Some required functions are met, but an important omission or factual error affects the response.	A usable sequence is present, but information is difficult to connect or retrieve in places.	Professional and unsuitable expressions are mixed enough to weaken the exchange.	Repeated errors require reader effort, although the main message is recoverable.
1	The response addresses only a small part of the request or makes a substantially misleading commitment.	Information is fragmentary or poorly sequenced, with little guidance for the recipient.	Tone or address is substantially unsuitable for the relationship and purpose.	Errors obscure much of the intended message; only fragments are reliably understood.
0	No required communicative function is meaningfully addressed.	No interpretable organization of the response is evident.	There is no interpretable attempt at a register appropriate to the communicative situation.	The written language does not permit a meaningful interpretation of the response.

A missing or unsubmitted script is recorded as missing, not automatically assigned zero. Scores describe evidence in an assessable response. Raters must justify a score using the response rather than the student's identity, effort, attendance, or condition.

by Tuesday at 16:00. These answers are available in the notes, so the exercise does not depend on inference about an unstated company policy.

Both conditions use the same model-identification exercise. Learners mark whether each model accurately addresses quantity, contents, delivery, price, and confirmation. They then identify a phrase that establishes a suitable professional relationship and a phrase that identifies the customer's next action. Model P1 supplies the relevant order facts and request for confirmation. Model P2 contains the deliberate quantity, delivery, and price problems identified in Appendix A. The exercise focuses on observable text features; it is not a standardized motivational questionnaire or a scored research endpoint.

B. Prompts That Define the Condition Difference

In the ARCS-informed orientation, use the prompt: "Name a situation in which you might need to reply to a customer, colleague, or service provider. Which part of this task would matter most in that situation?" A learner who has no current occupational example may choose an everyday service inquiry. The teacher accepts different purposes and connects the chosen example to accurate information and a clear next action. The comparison orientation states the same writing objective and asks learners to identify the topic and requested response from the packet.

During inquiry analysis, the ARCS-informed prompt is: "Which fact would change the customer's decision, and why?" The comparison explanation identifies the relevant facts, after which learners answer the common comprehension questions individually. During model comparison, the ARCS-informed condition asks learners to justify which formulation better serves the recipient. The comparison condition explains the

same differences and asks learners to label the corresponding features. Both conditions therefore encounter the content; the contrast concerns how its purpose is elicited and discussed.

Before the initial draft, the ARCS-informed planning prompt is: "Choose one communication requirement to check first. Write a short plan for meeting it." The comparison asks learners to begin the same response using the supplied criteria. At the end of the first lesson, the ARCS-informed reflection is: "Identify one choice you handled successfully and one specific change you will make next." The comparison closes with a factual review of the quantity, delivery arrangement, total, and deadline. Reflection is an opportunity to articulate progress, not a requirement to report enjoyment or confidence.

In the second lesson, both conditions apply feedback and produce an independent redraft of Practice Form P. The ARCS-informed condition revisits the learner's planned next step and asks for the communicative purpose of a revision. Its paired consolidation compares two alternative ways of explaining the same delivery or confirmation requirement. The comparison condition completes individual exercises using those same alternatives, then reviews the relevant language points. No additional assessment form is introduced during this consolidation. The final ARCS-informed reflection names a concrete improvement and a remaining difficulty; the comparison summarizes completed corrections.

These prompts are a reproducible starting script. A teacher may explain an unfamiliar word or restate a direction when needed, but should record substantial additional assistance or changes to the motivational activity. The record should identify what occurred rather than merely mark an overall lesson as compliant. If the comparison teacher spontaneously invites a learner-goal discussion or paired reflection, that overlap is documented as delivered, because it changes the observed

contrast even when the planned schedule remains unchanged.

C. Common Feedback and Language Exercises

Use three feedback categories in both conditions: missing requirement, inaccurate fact, and language that obstructs the intended message. An example of the first is “State when the customer must confirm.” An example of the second is “Check whether the stated total includes delivery.” An example of the third is “Make clear whether delivery is certain now or depends on confirmation.” Feedback points to the requirement and allows the learner to revise; it should not supply an entire replacement email for the learner to copy.

The common consolidation exercises ask learners to compare “We will send everything today” with “We can deliver the complete order before noon on Friday if you confirm by Tuesday at 16:00,” and “The price is 288 with delivery” with “The total, including delivery, is 296 currency units.” The first contrast concerns an unsupported commitment and a conditional arrangement; the second concerns a factual total. A further pair, “Tell us now” and “Please confirm by Tuesday at 16:00,” supports discussion of register and a usable deadline. These are practice contrasts, not additional observations or quantified evidence of learning.

The number and type of feedback comments should be recorded for each draft so that systematic differences in assistance can be examined. Identical feedback categories do not guarantee identical feedback amounts, because learners may produce different errors. The teacher should address the response using the common rules rather than force the same number of comments onto every script. In a study, analysis of feedback exposure should remain a process description unless a separate causal question and design justify further inference.

D. Independent Assessment Administration

Baseline administration occurs before the lesson sequence and before class allocation to the instructional conditions. Assign the A-then-B or B-then-A sequence using the prespecified balanced procedure, retaining the assignment even if a learner later misses an assessment. At baseline, give only the allocated form, its provider notes, and the common writing instructions. At primary follow-up, 24–48 hours after lesson two, give the other form. Administer Form C 14 days after lesson two within the prespecified three-day window on either side, recording the actual interval.

Use the same 20-minute limit and aid rules at every independent assessment. The default is no models, translation tools, generative assistance, peer help, or sentence-level teacher feedback. The administrator may clarify how to submit the response or how much time remains, without explaining what the learner should write. Keep the rubric accessible in the same way across conditions and occasions. If accessibility arrangements require a different duration, modality, or support, document and prospectively address those arrangements rather than silently classifying the responses as identical administrations.

Collect the response even when it is shorter than the approximate word range. Record whether a script was submitted and

whether it is assessable. A missing script, a corrupted file, and an assessable response containing few relevant functions are different states. Do not assign a zero total solely because a learner was absent or a file cannot be opened. Any effort to recover a file or clarify its submission status should precede scoring and be documented without exposing the learner’s condition to the raters.

Randomize the scoring order, remove identifiers that are unnecessary for rating, and provide raters with the actual inquiry and provider notes. Each rater records the four rubric scores, five content checks, and any material unsupported promise independently. The record retains the original ratings before discussion. For the composite, use the stated mean of the two independent totals. For factual completeness, retain both classifications and identify disagreement. All score corrections should distinguish a data-entry correction from a substantive rater reconsideration.

APPENDIX C. Plan for Empirical Examination

A. Content, Comprehension, and Administration

A feasibility study should begin with the intended use of the assessment: comparing performance on a bounded family of customer-response tasks under stated conditions. Reviewers with relevant language-teaching and occupational-communication experience can examine whether the supplied facts permit a defensible response, whether the rubric addresses the intended performance, and whether the tasks introduce avoidable background-knowledge demands. Their comments, disagreements, and resulting revisions should be recorded. Such review supplies evidence about content and interpretation; it does not establish an intervention effect or universal agreement about professional writing.

Learner comprehension can be examined with responses and short explanations from learners outside the subsequent effectiveness sample. Ask them to explain what the customer needs, what cannot be supplied exactly, and what decision their email requests. Observe where they return to the notes, which directions require clarification, and whether the time limit prevents a meaningful response. These observations should distinguish language difficulty from ambiguity in the materials. An unclear confirmation deadline, for example, should be repaired in the task rather than interpreted as a learner deficiency.

The content checks require explicit coding decisions. Each function receives 1 only when its specified elements are fulfilled accurately; a function requiring both total and deadline is incomplete when either is missing. A paraphrase can satisfy a function without copying the wording. Contradictory statements about the same arrangement do not count as accurate fulfillment merely because the correct fact appears elsewhere. A request for additional information is distinguished from a claim that an unavailable service is guaranteed. Raters record the relevant text when an omission, contradiction, or commitment determines the classification.

The study team should record completion time, requests for help, missing submissions, and the range of observed

response profiles for each form. No observed pilot percentage is prescribed here. Decision rules should be agreed before examining the feasibility results, including when factual ambiguity requires immediate revision, when an anchor requires clarification, and when a form must be replaced. Changes should be versioned so that subsequent estimates are linked to the materials actually used. A revised form is not automatically calibrated because its wording has become clearer.

B. Scoring Evidence and Consequences

Calibration should use scripts separate from masked outcome scoring. Raters first apply the rubric independently, compare the textual basis for their judgments, and discuss how the anchors handle factual accuracy, organization, register, and language control. Anchor revisions occur before trial scoring. Where the same fixed pair of raters scores all responses and inference concerns their average rating, the planned agreement analysis should explicitly identify a two-way mixed-effects, absolute-agreement, average-measures intraclass correlation for a suitable continuous total, following the distinctions described by Koo and Li [15]. Generalization to a population of raters would require a correspondingly justified rater design and model.

Report uncertainty in agreement estimates and the observed patterns of disagreement. A high correlation between raters can coexist with a systematic difference in score level, which is why the agreement definition matters. Dimension scores are ordinal and should also be examined using transparent cross-tabulations or another prespecified method suited to their interpretation. For the factual checks, report disagreements by function rather than treating agreement on one simple feature as agreement on the complete response. Do not tune the rubric repeatedly against the same masked outcome set and then report the resulting agreement as independent validation.

The intended score use should remain limited. The rubric supports classroom feedback and a prospective comparison of the specified task performance. It is not a credential for employment, a placement scale, or a CEFR certification. Examine whether equal weighting rewards behavior that conflicts with the communication objective, whether the approximate word range distracts from a concise answer, and whether adaptations change the response opportunity. Evidence adequate for a local formative use may be insufficient for a more consequential use.

C. From Feasibility to Comparative Evaluation

Before an effectiveness study, fix the material version, primary endpoint, form allocation, scoring rule, class allocation procedure, target estimand, and missing-data approach. Specify the educationally meaningful difference and planning assumptions using defensible external or feasibility evidence. The number of independently allocated classes, teacher arrangement, and anticipated cluster-size variation are substantive design choices. They should not be replaced by a large student count spread across only one class in each condition.

The comparative study should report both planned and actual delivery. A lesson that omits reflection, adds substantial

extra support, or exposes assessment forms during practice is a different implementation from the written plan. Retain the allocated-condition analysis while describing these departures; any supplementary analysis of actual exposure needs a separate interpretation because attendance and adherence are not randomly assigned. Report the direction, magnitude, and uncertainty of the primary contrast, together with dimension scores and factual-completeness descriptions. Mixed findings should be interpreted using the boundaries in Table 9, without changing the endpoint to obtain a favorable conclusion.

Data and Materials Availability

The data and materials supporting this article are available from the author upon reasonable request.

Use of Artificial Intelligence

Artificial intelligence tools assisted with correcting language mistakes, methodological drafting, LaTeX preparation, numerical checks, programming, reference checking, and figure preparation.

Conflicts of Interest

The author declares no conflicts of interest.

Funding

No specific funding was reported for this work.

References

- [1] Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing.
- [2] Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, 10(3), 2–10.
- [3] Li, K., & Keller, J. M. (2018). Use of the ARCS model in education: A literature review. *Computers & Education*, 122, 54–62.
- [4] Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32, 347–364.
- [5] Sun, L., & Asmawi, A. (2023). The effect of WeChat-based instruction on Chinese EFL undergraduates' business English writing performance. *International Journal of Instruction*, 16(1), 43–60.
- [6] Abdul Razak, R., Mat Yusoff, S., Hai Leng, C., & Mohamadd Marzaini, A. F. (2023). Evaluating teachers' pedagogical content knowledge in implementing classroom-based assessment: A case study among ESL secondary school teachers in Selangor, Malaysia. *PLOS ONE*, 18(12), Article e0293325.
- [7] Liu, G. L., Zhang, Y., & Zhang, R. (2024). Examining the relationships among motivation, informal digital learning of English, and foreign language enjoyment: An explanatory mixed-method study. *ReCALL*, 36(1), 72–88.
- [8] Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637.
- [9] Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, 10, 18–26.
- [10] Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A.-W., & Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ*, 348, Article g1687.
- [11] The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5–8.
- [12] Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- [13] Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- [14] Masters, K. (2013). Edgar Dale's Pyramid of Learning in medical education: A literature review. *Medical Teacher*, 35(11), e1584–e1593.

- [15] Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intra-class correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- [16] Alamer, A., Schuberth, F., & Henseler, J. (2024). When and how to use confirmatory composite analysis (CCA) in second language research. *Studies in Second Language Acquisition*, 46(2), 597–616.
- [17] Campbell, M. K., Piaggio, G., Elbourne, D. R., Altman, D. G., & CONSORT Group. (2012). CONSORT 2010 statement: Extension to cluster randomised trials. *BMJ*, 345, Article e5661.
- [18] Hopewell, S., Chan, A.-W., Collins, G. S., et al. (2025). CONSORT 2025 statement: Updated guideline for reporting randomised trials. *BMJ*, 389, Article e081123.
- [19] Kahan, B. C., Li, F., Copas, A. J., & Harhay, M. O. (2023). Estimands in cluster-randomized trials: Choosing analyses that answer the right question. *International Journal of Epidemiology*, 52(1), 107–118.
- [20] Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.

...