

AI Supported IELTS Preparation With Human Adherence Support Under Equal Planned Study Time

A Four Arm Stratified Randomized Trial

Article type: Original Research

Authors: Liwei Chen, Adaeze Okafor, Jonas Weber

Affiliation: External research team contracted under the Independence and Publication Agreement

Funding sponsor and product provider: Zhiying Technology (Hangzhou) Co., Ltd.

Abstract

Background: AI-supported language learning is increasingly offered alongside established forms of IELTS preparation, but randomized comparisons of complete preparation services remain limited.

Objective: To compare AI-supported digital preparation with non-instructional human adherence support (AI plus support) against self-study, large-group classes, and one-to-one tutoring under equal planned study time.

Methods: In this four-arm, individually randomized, stratified trial, 960 adults with estimated entry IELTS Academic bands of 4.5–6.0 were allocated equally to four preparation formats for 40 planned hours over 8 weeks. In the AI arm, instruction, practice, feedback, and question answering were delivered through the learning platform; human staff provided study supervision only. The primary outcome was a four-skill mock-test composite averaged across two post-test forms and adjusted for the average of two pre-test forms. Writing and speaking were independently scored by masked raters. The intention-to-treat analysis used multiple imputation and accounted for clustering within classes, tutors, and adherence-support staff caseloads. Three prespecified comparisons used Bonferroni-adjusted 98.33% confidence intervals, providing at least 95% simultaneous coverage. The primary criterion required all three lower bounds to exceed zero.

Results: Post-test completion was 82.4% (791/960). Adjusted differences favored AI plus support over self-study (0.54 composite-score units, 98.33% CI 0.33–0.76), large-group classes (0.45, 0.20–0.71), and one-to-one tutoring (0.45, 0.21–0.69); all Bonferroni-adjusted p values were < .001. Complete-case estimates were similar. Satisfaction estimates favored AI plus support over self-study and large-group classes; the comparison with one-to-one tutoring did not establish superiority or equivalence.

Conclusions: Under equal planned study time, assignment to AI plus support improved mock-test performance relative to the three comparison services as implemented in this trial. The findings concern the combined service and do not establish AI-specific effects, per-hour efficiency, or gains in official IELTS scores.

Keywords: IELTS; AI-assisted language learning; intelligent tutoring systems; human adherence support; randomized controlled trial; test preparation

1 Introduction

1.1 IELTS preparation and language learning

IELTS Academic preparation combines language development with learning to perform under the constraints of a particular assessment. Candidates must allocate limited time across listening, reading, writing, and speaking while choosing among materials and services that differ in instructional support, feedback, and structure. Understanding the effects of these choices is relevant to language education as well as to candidates making practical decisions about preparation.

Previous research cautions against treating preparation time as a reliable predictor of score improvement. Elder and O'Loughlin (2003) documented substantial variation in IELTS gains during intensive English study. Research on washback also shows why test preparation should be examined in relation to both instructional practices and the outcomes being assessed (Alderson & Wall, 1993; Green, 2007a, 2007b). Better performance on practice tests can be useful to learners, but it does not by itself establish broader language development, durable learning, or improved performance on an official examination.

Comparisons of preparation formats are complicated by differences in available study time, teacher support, learner preferences, and assessment conditions. A comparison under a common planned time budget can address one practical question: what happens when learners are assigned different preparation arrangements with the same expected time commitment? Such a design estimates the effect of assignment to a service configuration. It does not hold actual engagement constant or establish the relative efficiency of each hour studied.

1.2 Technology and human support

Intelligent tutoring systems provide a rationale for examining technology-supported preparation, although evidence from other subjects and educational settings cannot establish the effectiveness of a particular language-learning product. Reviews of intelligent tutoring systems describe learning benefits that vary with the instructional context, comparison condition, and outcome measure (Kulik & Fletcher, 2016; Ma et al., 2014; VanLehn, 2011). These findings motivate product-specific evaluations using appropriate comparison groups and assessments that are not generated or scored by the product itself.

Recent meta-analyses of chatbot-assisted language learning also report average benefits across comparative studies (Lyu et al., 2025; Wang et al., 2025). These findings support further evaluation while leaving the effects of specific products and service configurations to be tested.

Human support may also matter when instruction is delivered through a digital platform. Self-regulated learning involves planning, monitoring progress, and sustaining engagement (Zimmerman, 2000). A learner may therefore benefit from access to instructional resources while still requiring support to maintain study. In the service evaluated here, all learning activities, practice, instructional feedback, and question answering were delivered through the platform. Human staff supervised study adherence through check-ins and reminders. They did not teach or

answer subject-matter questions. The human contribution was thus non-instructional adherence support.

This combination raises two distinct questions. The first concerns the effectiveness of the complete service that learners receive. The second concerns the contributions of its components and the processes through which they operate. A comparison of an AI-supported service with other preparation formats can address the first question. The present design does not separate the effects of platform-delivered learning from those of non-instructional adherence support, their interaction, or other differences between the implemented services. A comparison of the same platform with and without adherence support would address the additional contribution of the human component.

1.3 The present study

The present four-arm randomized trial compared the Yingwu Zhixue platform with non-instructional human adherence support (AI plus support) against self-study, large-group instruction, and one-to-one tutoring. Participants were adults preparing for IELTS Academic with estimated entry levels of 4.5–6.0. Each arm had a planned study budget of 40 hours over eight weeks, including instruction, assignments, and individual practice. The comparisons concern the particular services delivered in this trial and do not represent every provider or implementation of these formats.

The primary outcome was a four-skill standardized mock-test composite, assessed using two post-test forms and adjusted for the average of two pre-test forms. Writing and speaking were independently scored under masked conditions. The protocol and statistical analysis plan were registered before recruitment. The three prespecified hypotheses were that assignment to AI plus support would produce higher post-test composite performance than assignment to self-study, large-group instruction, and one-to-one tutoring, respectively. The prespecified overall criterion required all three comparisons to favor AI plus support with all three Bonferroni-adjusted confidence intervals lying above zero.

Skill-specific outcomes, satisfaction, and actual study time provided additional perspectives on the intervention experience. These outcomes were interpreted separately from the primary comparisons. In particular, satisfaction was assessed in participants who knew their allocation, and actual study time was a post-randomization measure. The study therefore evaluates a service package over a defined preparation period, without attributing effects to AI alone or translating the composite into official IELTS band gains.

2 Methods

2.1 Study design and preregistration

This was a four-arm, parallel-group trial with individual randomization in a 1:1:1:1 ratio, stratified by entry proficiency and recruitment batch. Participants shared delivery units in three arms: instructional classes, individual tutors, and adherence-support staff caseloads. The human staff in the AI arm supervised study adherence and did not provide instruction. Self-study

participants did not share an instructional unit. The design therefore required an analysis that accommodated partial clustering (Baldwin et al., 2011).

The protocol and statistical analysis plan were registered before recruitment. Reporting was guided by CONSORT 2025, with protocol elements aligned with SPIRIT 2025 and analyses informed by methodological guidance for multi-arm trials (Chan et al., 2025; Hopewell et al., 2025a, 2025b; Odutayo et al., 2020). Participants and service providers knew their allocations. Outcome raters and the primary analysts were masked to treatment assignment.

2.2 Participants and recruitment

Eligible participants were adults aged 18 years or older preparing for IELTS Academic, with an estimated entry band of 4.5–6.0 on the study's calibrated pre-test forms. They had to be able to attend standardized testing sessions, accept assignment to any of the four formats, and refrain from concurrent enrollment in conflicting preparation programs. The screening estimate was used to determine eligibility and stratification; it was not an official IELTS score.

Exclusion criteria were duplicate registration, participation in the measurement pilot, known prior exposure to the designated post-test forms that could not be addressed by form reassignment, and inability to provide informed consent. Previous official IELTS results were recorded descriptively and did not replace study-administered baseline testing.

Recruitment used public advertising across multiple channels rather than recruitment restricted to existing product users or fan communities. Before randomization, participants reported their format preferences, expectations of effectiveness, recent preparation history, and available study time. The resulting sample represents candidates willing to accept random allocation across these preparation formats.

2.3 Randomization and allocation concealment

Participants completed consent procedures and two pre-test forms, A and B, at least 48 hours apart before allocation. An independent statistician generated the allocation sequence using variable block sizes within proficiency strata and recruitment batches. The two proficiency strata corresponded approximately to entry bands of 4.5–5.0 and 5.5–6.0.

Recruiters, teachers, support staff, and sales staff did not have access to the allocation sequence. The research system retained allocation timestamps, operator records, and audit trails.

2.4 Interventions and planned study time

Each arm was assigned 40 hours of study over 8 weeks, equivalent to 5 hours per week. This budget included instruction, platform use, assignments, independent practice across the four skills, and learner-initiated instructional contact. Research tests, questionnaires, and administrative contact were excluded. Support work performed without the learner was recorded as service labor rather than learner study time. The 40-hour allocation was a total planned study budget, rather than 40 hours of teacher contact in each arm.

The self-study arm received a complete licensed resource pack containing authentic-format practice materials, audio, answer keys, and published band descriptors. Participants directed their own practice and did not receive individualized teacher feedback or human adherence support. Resource eligibility was determined before the trial.

The large-group arm attended a complete, normally operating short course appropriate to the target entry level, including its usual assignments, feedback, and opportunities for questions. Participants joined courses at their normal starting point. Fourteen teaching units contributed study participants; actual class sizes were 30–60 students, with only a subset enrolled in the trial. Students outside the trial were not included in the dataset.

The one-to-one arm received individual tutoring with the assignments and feedback provided in the commercial program. Eighteen tutors delivered the intervention. Tutor qualification requirements were specified before the trial, and the product team did not select the tutors.

The AI plus support arm used Yingwu Zhixue (Parrot Learning), a commercial AI-supported learning platform comprising a mobile app and a companion web interface. Participants had access to personalized instruction, practice, automated feedback, an AI study assistant, and mock examinations throughout the eight-week intervention. The following description concerns the IELTS Academic functions available during the trial; availability does not imply that every participant used every function.

The app generated individualized lesson plans and instructional content using the learner's target score, current proficiency, and identified areas of difficulty. A real-time voice interface allowed learners to interrupt explanations, ask follow-up questions, and request alternative examples. Performance information from mock tests, practice activities, spoken responses, writing, and vocabulary tasks informed an individual learning profile. The system updated this profile and recalculated the next day's study plan, linking feedback from completed activities to the selection of subsequent instruction, practice, and review.

Targeted practice addressed specific difficulties across the four skills. Speaking activities included word- and sentence-level pronunciation practice, shadowing, and responses to IELTS Speaking Parts 1–3. Writing activities included sentence revision, paragraph practice, and feedback on Tasks 1 and 2. Listening activities included dictation, keyword recognition, and practice with connected speech; reading activities addressed complex sentences, paraphrase recognition, and locating relevant information. Vocabulary activities included recognition of spoken words, spelling, and pronunciation. Tasks could progress from individual words and sentences to paragraphs and complete responses, with repeated practice selected in relation to the learner's profile.

Automated feedback combined suggested revisions with explanations of identified difficulties. Speaking feedback addressed pronunciation, stress, intonation, fluency, vocabulary, and grammar. Writing feedback addressed task fulfillment, organization, vocabulary, and grammar at sentence and paragraph levels. The relevant IELTS assessment criteria informed this feedback (IELTS, 2026a). Estimated band scores and diagnostic reports were used for formative learning,

and feedback informed subsequent updates to the learning profile and study plan. These outputs were not official IELTS scores and did not determine the trial's primary outcome scores.

The AI study assistant, Yingzai, supported task scheduling, access to practice materials and feedback reports, review reminders, progress summaries, and answers to basic IELTS questions. The companion web interface provided timed mock examinations covering listening, reading, writing, and speaking. Objective scoring for listening and reading and AI-generated estimates and feedback for writing and speaking produced reports that informed subsequent practice. Platform mock examinations were learning activities, distinct from the independently administered research assessments described in Section 2.5.

Human staff provided non-instructional adherence support through study check-ins and reminders. They did not teach, provide instructional feedback, or answer subject-matter questions; these learning functions were delivered through the platform. Human adherence support was organized across 16 staff caseloads. Product versions and changes during the trial were logged. Additional human instruction, completion of practice on participants' behalf, and rewards tied to scores were prohibited. The intervention therefore combined the platform's instructional and automated study-support functions with human supervision of study adherence.

The four formats differed in how learning activities were delivered and in the role of human staff (Table 1). These descriptions refer to the service configurations implemented in this trial.

Table 1 Learning activities and human roles across the four preparation formats

Preparation format	Learning activities and feedback	Human role
Self-study	Self-directed practice using a licensed resource pack, audio, answer keys, and band descriptors.	No individualized teacher feedback or adherence support.
Large-group classes	Group lessons with the course's usual assignments, feedback, and opportunities for questions.	Teachers delivered the course and its usual instructional support.
One-to-one tutoring	Individual lessons with the commercial program's assignments and feedback.	Tutors provided individual instruction and feedback.
AI plus support	Personalized voice instruction, targeted practice, automated feedback, and an AI study assistant in the app; timed mock tests through the companion web interface.	Staff supervised study adherence through check-ins and reminders; they did not teach or answer subject-matter questions.

All four arms had the same planned total study budget of 40 hours over 8 weeks. The AI arm combined platform-delivered learning with human adherence support; the trial did not isolate the contribution of either component.

Weekly logs used common definitions to record actual study time, preparation outside the assigned program, tool use, human contact, and deviations. Rules for platform-derived active

learning time were evaluated during the pilot. Research contact was restricted to scheduling, log completion, equipment, and safety matters. Motivational reminders formed part of each service's intervention rather than additional research support. Compensation was not contingent on scores, satisfaction, or task completion.

2.5 Assessment and outcome construction

Four complete authentic-format tests were selected by the independent assessment team from licensed Cambridge IELTS Academic practice materials. Forms A and B were administered before the intervention, and forms C and D afterward. Post-test forms were not used in baseline assessment or instruction. Practice-test results were interpreted as study outcomes rather than official IELTS scores (Cambridge University Press & Assessment, 2026).

Within each assessment stage, participants across arms took the same forms within a common testing window. Form order was balanced within proficiency strata and recruitment batches. Testing sessions included participants from different arms. Timing, environment, equipment, and response mode were standardized.

Listening and reading were scored using frozen answer keys and tolerance rules. Item responses and raw scores on the 0–40 scales were retained. Automated checks and sampled double entry supported verification, and score edits were logged with the original value, reason, operator, and timestamp. Generic raw-score conversion tables were not used as substitutes for form-specific equating (IELTS, 2026a).

Writing and speaking were independently scored by two raters trained on the public IELTS band descriptors and calibrated with study anchor samples. Writing followed the Task 1 and Task 2 criteria, with Task 2 weighted twice as heavily. Speaking used four equally weighted criteria. Speaking interviews followed a common protocol and were recorded (IELTS, 2026a, 2026b).

Scripts and recordings were stripped of participant identifiers, treatment labels, assessment-stage labels, and service-identifying cues. Baseline and post-test materials were mixed for scoring. Raters were masked to each other's scores, baseline performance, and the study hypotheses; where feasible, a rater did not assess the same participant twice. Skill-score disagreements exceeding 0.5 band were referred to a third independent rater, with the median of the three scores used as the final score. Otherwise, the two scores were averaged. These rules were fixed before formal scoring. Product-generated AI scores were not used as primary outcome scores.

Anchor or duplicate materials constituted approximately 10% of each scoring batch. Quality monitoring used the absolute-agreement intraclass correlation coefficient (ICC) for the mean of two raters, alongside disagreement and adjudication rates (Shrout & Fleiss, 1979). The study quality target was $ICC \geq .80$; this was a project criterion rather than an official IELTS standard. Batches that failed quality checks were investigated and rescored following retraining before treatment labels were revealed.

An independent measurement pilot of approximately 80 participants, excluded from the randomized trial, established centering and scaling constants for each skill on each form. Each

form's four standardized skill scores were averaged with equal weights. The two form-level composites were then averaged within each assessment stage. The constants, scoring code, missing-item rules, and handling of extreme values were fixed before recruitment.

The primary endpoint was the post-test composite, with the baseline composite used as a covariate. Effects are reported in units of this composite. Averaging standardized skill scores does not imply that the resulting composite has a standard deviation of one. Because forms were not equated to a common official band scale, composite differences were not converted into official IELTS band gains. Interpretation focused on randomized between-arm differences at post-test rather than absolute changes across different forms.

2.6 Secondary and exploratory outcomes

Secondary outcomes included listening and reading raw scores, writing and speaking study scores, overall satisfaction, post-test completion, withdrawal, actual study time, additional preparation, and service contact. Skill-level outcomes were interpreted as secondary estimates rather than additional confirmatory tests of the primary hypothesis. Exploratory analyses examined variation by entry stratum and associations between study time and outcomes. Cost-related measures were designated exploratory; no cost-effectiveness claim is made in this report.

The key secondary outcome was overall satisfaction, measured after the course and before post-testing. Participants answered, “Considering the whole 8 weeks, how satisfied are you overall with the study arrangement you were assigned?” on a scale from 0, extremely dissatisfied, to 10, extremely satisfied. Auxiliary agreement items used separate 1–7 scales. Cognitive interviews during the pilot assessed comprehension across arms. The questionnaire was designed for this study and was not presented as a validated standardized instrument. Because allocation was known to participants, satisfaction remained susceptible to preferences and expectations.

2.7 Sample size considerations

Planning addressed the joint objective of superiority over all three comparison arms. The initial calculations assumed a minimum standardized difference of $d = 0.50$, defined relative to the marginal residual standard deviation inclusive of clustering, with three two-sided Bonferroni-adjusted tests. Approximately 93% power per comparison was used to target approximately 80% joint power. Calculations allowed for 20% attrition and partial clustering, with approximate unit sizes of 20 for classes and 15 for tutor and adherence-support staff caseloads.

Normal-approximation calculations suggested approximately 812 randomized participants under $ICC = .02$, increasing to approximately 1,112 under $ICC = .05$. A smaller assumed difference of $d = 0.40$ under the moderate-clustering scenario required approximately 1,268 participants. The trial recruited 960 participants, 240 per arm, as a capacity- and procurement-constrained target. These scenarios describe the planning trade-offs; they do not establish that the achieved design had at least 80% joint power under every plausible effect size or clustering configuration. The number and size of delivery units were relevant to both precision and power (U.S. National Institutes of Health, 2026).

2.8 Statistical analysis

The primary analysis followed the intention-to-treat principle, retaining participants in their assigned arms regardless of adherence. The post-test composite was regressed on treatment arm, the baseline composite, the stratification factors, recruitment batch, and prespecified form-order terms. Mixed models accommodated arm-specific clustering within classes, tutors, and adherence-support staff caseloads. Cluster-robust variance estimation was prespecified as a fallback and robustness approach for the limited number of delivery units (Baldwin et al., 2011). Analyses that treated all participants as independent without accounting for the shared delivery units were excluded from the primary strategy.

The three confirmatory contrasts compared AI plus support with self-study, large-group classes, and one-to-one tutoring. Each contrast used a two-sided 98.33% confidence interval and a Bonferroni-adjusted p value. Taken together, these intervals provide at least 95% simultaneous coverage across the three comparisons (Bender & Lange, 2001). The prespecified superiority criterion required all three lower bounds to exceed zero. Statistical superiority was interpreted separately from educational importance; failure to detect a difference was not treated as evidence of equivalence or non-inferiority.

Participants were invited to complete post-testing even if they discontinued study activities. The primary approach to missing outcomes used 20 multiply imputed datasets under a missing-at-random assumption, preserving treatment arm, baseline performance, recruitment batch, and cluster structure (Rubin, 1987). A blinded committee adjudicated suspected test leakage or misconduct. Invalidated outcomes were treated as missing while the participants remained in the intention-to-treat population.

Prespecified sensitivity analyses included a complete-case analysis and departures from the missing-at-random assumption. In the reported shift analysis, imputed post-test values in the AI plus support arm were reduced relative to their missing-at-random values. A per-protocol analysis used a minimum of 28 recorded study hours. Because adherence is a post-randomization characteristic, per-protocol estimates were treated as potentially affected by selection bias and were not used to establish robustness to missing outcomes.

Scoring quality checks, cleaning, anomaly adjudication, and statistical analysis plan sign-off preceded unmasking. Analysts rehearsed the analysis code with masked arm labels, and the database was locked before the final analysis.

2.9 Ethics and research independence

The study received independent institutional ethics approval before recruitment. Informed consent covered random allocation, the four preparation services, the mock-test nature of the outcomes, recording and logging, compensation, withdrawal, and data handling. Contact information was stored separately from analysis data. Access, retention, withdrawal, and destruction arrangements were specified in the consent materials.

An Independence and Publication Agreement was signed before recruitment. The sponsor funded the study and delivered the product arm under routine service conditions. The external research team controlled the protocol, recruitment integrity, randomization, assessment, analysis, and publication. The sponsor could review factual accuracy and confidential information but could not suppress findings or determine whether results were published. Test-material permissions were documented, and study administration used licensed materials (Cambridge University Press & Assessment, n.d.; IELTS, 2026c).

3 Results

3.1 Participant flow and study time

Between 11 May and 5 June 2026, 1,466 candidates were assessed for eligibility. Of the 506 excluded candidates, 158 were outside the entry-band range, 214 declined randomization, 77 could not attend the required sessions, 56 had prior exposure to designated forms that could not be addressed by reassignment, and one had registered twice. The remaining 960 participants were randomized, with 240 allocated to each arm.

Overall, 791 participants completed post-testing and 169 did not, corresponding to completion and missingness rates of 82.4% and 17.6%, respectively. The primary analysis included all 960 randomized participants through multiple imputation. Completion counts and recorded study time are presented in Table 2.

Table 2 Participant completion and recorded study time

Measure	AI plus support	Large-group classes	Self-study	One-to-one tutoring
Randomized participants	240	240	240	240
Completed post-testing	195	196	199	201
Missing post-test outcome	45	44	41	39
Actual study time in hours, M (SD)	37.4 (5.3)	38.7 (4.6)	29.6 (9.3)	38.8 (4.0)

All groups were assigned 40 planned study hours. Actual study time differed across arms, most notably in self-study, where the mean was 29.6 hours, or 74.0% of the planned amount. Recorded study time was summarized descriptively and was not used to adjust the primary treatment comparison, because it was measured after randomization and could itself be affected by the assigned service.

3.2 Baseline characteristics

Baseline characteristics are summarized in Table 3. Mean age was approximately 24 years in each arm, and baseline composite means ranged from 0.00 to 0.08.

Table 3 Baseline characteristics by randomized arm

Characteristic	AI plus support	Large-group classes	Self-study	One-to-one tutoring
Participants, n	240	240	240	240
Age in years, M (SD)	23.7 (4.0)	23.6 (3.9)	23.8 (3.9)	23.7 (3.7)
Female, %	60.4	56.2	65.0	59.2
Higher entry stratum, %	55.0	51.2	55.0	57.9
Previous official IELTS sitting, %	36.2	36.7	38.3	41.7
Baseline composite, M (SD)	0.03 (0.76)	0.00 (0.78)	0.05 (0.75)	0.08 (0.77)

The higher entry stratum corresponded approximately to estimated bands of 5.5–6.0. Composite values are expressed in the study-defined units described in Section 2.5.

3.3 Scoring quality

The absolute-agreement ICC for the mean of two raters was .87 for writing and .83 for speaking, both above the study target of .80. The mean absolute difference between writing raters was 0.45 band. Third-rater adjudication was required for 20.0% of writing scripts and 27.1% of speaking assessments. One early scoring batch prompted retraining; subsequent anchor checks passed. Agreement statistics were calculated from scores recorded before adjudication.

3.4 Primary outcome

Adjusted post-test composite means were 0.91 for AI plus support, 0.45 for large-group classes, 0.36 for self-study, and 0.46 for one-to-one tutoring. Table 4 presents the primary intention-to-treat contrasts after multiple imputation and adjustment for baseline performance and clustering.

Table 4 Primary adjusted contrasts in post-test composite scores

Contrast	Difference	SE	98.33% CI	Adjusted p
AI plus support minus self-study	0.54	0.09	0.33 to 0.76	< .001
AI plus support minus large-group classes	0.45	0.11	0.20 to 0.71	< .001
AI plus support minus one-to-one tutoring	0.45	0.10	0.21 to 0.69	< .001

Differences and standard errors are in composite-score units. Confidence intervals and p values incorporate Bonferroni adjustment for the three confirmatory contrasts; the intervals provide at least 95% simultaneous coverage. Arm means and contrast estimates were rounded separately.

All three lower confidence bounds exceeded zero, meeting the prespecified criterion of superiority over each comparison arm. Estimated cluster ICCs were .08 for large-group classes, .07 for one-to-one tutoring, and .02 for adherence-support staff caseloads. The class and tutor

ICC estimates exceeded the .02 and .05 scenarios summarized in the initial sample-size calculations, reinforcing the importance of accounting for clustering in the analysis.

3.5 Skill outcomes

Secondary estimates favored AI plus support across listening, reading, writing, and speaking (Table 5). These results describe the pattern across skills and do not constitute additional multiplicity-controlled confirmatory claims.

Table 5 Adjusted skill differences for AI plus support versus each comparison arm

Skill	Versus self-study	Versus large-group classes	Versus one-to-one tutoring
Listening raw score, 0–40	3.05	2.52	2.38
Reading raw score, 0–40	2.62	2.14	2.16
Writing study score	0.45	0.44	0.40
Speaking study score	0.46	0.40	0.39

Values are adjusted point estimates. These secondary comparisons are presented descriptively and do not support separate claims of superiority for each skill. Writing and speaking study scores use the public band-descriptor scales; they are not official IELTS scores.

3.6 Satisfaction

Among completers, mean overall satisfaction was 7.98 for AI plus support, 7.02 for large-group classes, 6.68 for self-study, and 7.74 for one-to-one tutoring. Adjusted differences for AI plus support were 1.29 points versus self-study (95% CI 0.98–1.61), 0.97 versus large-group classes (0.64–1.29), and 0.23 versus one-to-one tutoring (–0.08 to 0.54).

The secondary estimates favored AI plus support over self-study and large-group classes. The interval for the one-to-one comparison included zero; this analysis established neither superiority nor equivalence. These 95% intervals should not be interpreted as having the simultaneous coverage used for the primary outcome. Satisfaction also remained subject to expectation effects and selection associated with survey completion.

3.7 Sensitivity analyses

In the complete-case analysis of 791 participants, adjusted differences for AI plus support were 0.54 versus self-study, 0.45 versus large-group classes, and 0.44 versus one-to-one tutoring. These point estimates were close to the corresponding primary estimates.

The per-protocol subset defined by at least 28 recorded study hours comprised 776 participants, including 122 in the self-study arm. The corresponding adjusted point estimates were 0.56, 0.45, and 0.45, respectively. This adherence-defined sample is distinct from the complete-case sample.

Its estimates were considered supplementary because conditioning on adherence can introduce selection bias; they were not used to rule out bias from missing outcomes.

In the missing-not-at-random sensitivity analysis, imputed post-test values in the AI plus support arm were shifted downward while the comparison arms were left unchanged. At a downward shift of 0.6 residual standard deviations, the lower 98.33% confidence bound for AI plus support versus one-to-one tutoring remained positive at 0.10. This result describes robustness under the specified shift scenario. Similar completion rates across arms do not establish that missingness mechanisms were similar, and this analysis does not exclude bias under other departures from the missing-at-random assumption.

3.8 Exploratory analysis by entry stratum

The arm-by-stratum analysis yielded an interaction estimate of -0.34 (95% CI -0.65 to -0.03) for the class contrast. This exploratory estimate was not used to establish superiority within either entry stratum. Interpretation requires caution because multiple interactions were examined and stratum-specific superiority was not the primary objective.

4 Discussion

4.1 Principal findings

Under a common planned study budget of 40 hours over eight weeks, assignment to AI-supported digital preparation with non-instructional human adherence support resulted in higher adjusted mock-test composite performance than assignment to each of the three comparison services. The estimated differences were 0.54 composite-score units relative to self-study, 0.45 relative to large-group instruction, and 0.45 relative to one-to-one tutoring. All three Bonferroni-adjusted confidence intervals lay above zero, meeting the prespecified overall criterion.

These results support the effectiveness of the implemented AI-plus-support package for the population and assessment conditions studied. They do not establish superiority over all self-study resources, classes, or tutors. Composite-score differences also cannot be interpreted directly as official IELTS band gains or as a common standardized effect size across studies. Their practical significance requires consideration of the outcome scale, its uncertainty, and the decisions learners are making.

4.2 Interpretation for language learning

The secondary skill-level estimates favored AI plus support across listening, reading, writing, and speaking. This pattern suggests that the primary composite difference was not confined to a single skill. However, the skill-level analyses remain secondary and should be read with their multiplicity and uncertainty in view. They do not establish separate confirmatory claims for each language skill.

The findings extend research on technology-supported learning by examining an implemented service against several preparation arrangements within the same randomized study. Previous reviews provide relevant context (Kulik & Fletcher, 2016; Ma et al., 2014; VanLehn, 2011), but

differences in populations, content, comparator quality, and outcome construction prevent direct numerical comparisons. The present results concern performance on independently administered, authentic-format practice tests. They leave open whether the observed differences persist, transfer to less test-specific language tasks, or predict performance on an official IELTS examination.

The platform linked performance information to subsequent lesson planning, targeted practice, and review. Interactive explanations, repeated practice with feedback, and revisions to daily study plans could plausibly contribute to learning, while automated reminders and human check-ins could support engagement. The trial did not isolate these processes or establish the contribution of any individual feature. Although human staff did not provide instruction, their supervision could have affected adherence. Whether adherence support contributed to the observed advantage remains a hypothesis for subsequent research.

4.3 Study time and satisfaction

Actual study time differed between arms despite the common plan. Self-study participants logged an average of 29.6 hours, equivalent to 74% of the planned budget, compared with 37.4 hours in AI plus support, 38.7 hours in large-group instruction, and 38.8 hours in one-to-one tutoring. These descriptive differences are relevant to the effect of assigning a preparation format: the support available within a service may affect whether learners carry out their plans.

Nevertheless, study time was measured after randomization and may reflect motivation, difficulty, progress, and other learner characteristics. The results do not show how much of the outcome difference was mediated by engagement, nor do they establish performance per hour. Adjusting the primary comparison for realized hours would address a different question and could introduce bias. Future work could combine more detailed process measurement with designs intended to test specific engagement mechanisms.

Satisfaction favored AI plus support over self-study and large-group instruction. The estimated difference from one-to-one tutoring was 0.23 points on the 0–10 scale, with a 95% confidence interval from –0.08 to 0.54. This comparison did not establish equivalence. Satisfaction also reflects expectations and preferences because participants knew their assignment, and the reported completer analysis may not represent all randomized participants.

4.4 Strengths and limitations

Random allocation, a common planned time budget, repeated assessment forms, and independent masked scoring strengthen the interpretation of the between-group comparisons. Intention-to-treat analysis addresses the effect of assignment, including differences in adherence. The sponsor's role was disclosed, with responsibilities for allocation, assessment, analysis, and publication assigned to an external research team under an independence agreement. These arrangements support scrutiny of the findings but do not eliminate the need for independent replication.

Post-test data were available for 82.4% of randomized participants. Multiple imputation relied on a missing-at-random assumption, and the sensitivity analyses examined specified departures from that assumption. Similar completion rates across arms do not establish similar reasons for missingness or rule out informative missing data. The reported sensitivity results should therefore be understood within the scenarios examined.

Other limitations concern the product configuration and market context studied, participants' willingness to accept random allocation, and the number of teaching and adherence-support units. The eight-week endpoint does not address retention. Practice-test exposure, familiarity with assessment demands, and novelty may also influence performance. The analyses reported here did not validate the accuracy of automated feedback or estimated band scores, or assess the effects of exposure to individual platform functions. Finally, the combined intervention prevents attribution to individual components, and the results reported here do not establish cost-effectiveness.

4.5 Implications and conclusion

For language-education research, the next steps are replication across service providers and learner populations, follow-up assessment, and comparisons of the same platform with and without non-instructional adherence support. Broader language tasks and linkage to official examination outcomes would help establish the scope and practical relevance of the observed differences.

For adults resembling those enrolled in this trial, the findings provide evidence that the studied AI-plus-support service can improve mock-test performance relative to the three implemented alternatives under the same planned time commitment. This conclusion concerns the complete service, the specified preparation period, and the study's outcome measure. It does not imply equivalence in satisfaction, an AI-specific mechanism, or an official IELTS score increase.

Statements

Ethics approval and consent: The protocol was approved by an independent institutional review board before recruitment. All participants provided informed consent covering randomization, recording, logging, compensation, withdrawal, and data handling.

Preregistration: The protocol and statistical analysis plan were registered before recruitment. The measurement pilot and the main randomized trial were registered separately.

Data availability: De-identified analysis data, a data dictionary, and randomization, analysis, and simulation code are available from the research team upon reasonable request, subject to participant privacy and test-material licensing restrictions. Protected test content is excluded from data sharing.

Funding: The study was funded by Zhiying Technology (Hangzhou) Co., Ltd., which also provides the product evaluated in the AI plus support arm. An Independence and Publication Agreement signed before recruitment restricted the sponsor's role in randomization, assessment,

scoring, analysis, and publication decisions and guaranteed publication of results regardless of direction.

Competing interests: The authors report no financial relationship with the sponsor beyond the research contract described above.

Author contributions: Conceptualization: Chen, Okafor, and Weber. Methodology: Chen, Okafor, and Weber. Investigation: Chen and Okafor. Formal analysis: Weber. Writing of the original draft: Chen. Review and editing: all authors.

Acknowledgements: We thank the independent raters, interviewers, and proctors, and all 960 participants enrolled in the trial.

References

- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129.
- Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods*, 16(2), 149–165.
- Bender, R., & Lange, S. (2001). Adjusting for multiple testing — when and how? *Journal of Clinical Epidemiology*, 54(4), 343–349.
- Cambridge University Press & Assessment. (2026). *Cambridge IELTS 20 Academic — official introduction and practice test notes*.
https://assets.cambridge.org/97810098/14904/excerpt/9781009814904_excerpt.pdf
- Cambridge University Press & Assessment. (n.d.). *Rights and permissions*.
<https://www.cambridge.org/pg/rights-and-permissions>
- Chan, A.-W., Boutron, I., Hopewell, S., Moher, D., Schulz, K. F., Collins, G. S., et al. (2025). SPIRIT 2025 statement: Updated guideline for protocols of randomised trials. *BMJ*, 389, e081477.
- Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive English language study and band score gain on IELTS. *IELTS Research Reports*, 4, 207–254.
- Green, A. (2007a). *IELTS washback in context: Preparation for academic writing in higher education* (Studies in Language Testing 25). Cambridge University Press and Cambridge ESOL.
- Green, A. (2007b). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education*, 14(1), 75–97.
- Hopewell, S., Chan, A.-W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., et al. (2025a). CONSORT 2025 statement: Updated guideline for reporting randomised trials. *BMJ*, 389, e081123.
- Hopewell, S., Chan, A.-W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., et al. (2025b). CONSORT 2025 explanation and elaboration: Updated guideline for reporting randomised trials. *BMJ*, 389, e081124.
- IELTS. (2026a). *IELTS scoring in detail*. <https://ielts.org/take-a-test/your-results/ielts-scoring-in-detail>
- IELTS. (2026b). *Resources for setting your IELTS scores*. <https://ielts.org/organisations/ielts-for-organisations/understanding-ielts-scoring/resources-for-setting-your-ielts-scores>
- IELTS. (2026c). *IELTS copyright and trade mark statement*. <https://ielts.org/legal/ielts-copyright-and-trade-mark-statement>
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78.
- Lyu, B., Lai, C., & Guo, J. (2025). Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies. *International Journal of Applied Linguistics*, 35(2), 834–851.
- Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918.
- Odutayo, A., Gryaznov, D., Copsey, B., Monk, P., Speich, B., Roberts, C., et al. (2020). Design, analysis and reporting of multi-arm trials and strategies to address multiple testing. *International Journal of Epidemiology*, 49(3), 968–978.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.
- U.S. National Institutes of Health. (2026). *Research methods resources: Individually randomized group-treatment trials*. <https://researchmethodsresources.nih.gov/methods/irgt>
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221.
- Wang, F., Cheung, A. C. K., Neitzel, A. J., & Chai, C. S. (2025). Does chatting with chatbots improve language learning performance? A meta-analysis of chatbot-assisted language learning. *Review of Educational Research*. Advance online publication. <https://doi.org/10.3102/00346543241255621>
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press.