

Mobile-based English language learning through bite-sized formative CEFR- aligned assessment

Michael Milanovic
Catherine Jones

May 2026

Mobile-based English language learning through bite-sized formative CEFR-aligned assessment

Michael Milanovic and Catherine Jones

Abstract

This paper reports on a large-scale pilot implementation of CATs Core, a mobile English language learning and assessment system covering CEFR levels A0 to B1. Conducted in operational educational contexts, the study focuses on usability, engagement, and indicative learning outcomes in real-world settings, rather than on controlled experimental validation. The analyses are therefore descriptive and exploratory, providing a foundation for subsequent, more rigorous validation research. CATs Core forms part of the wider CATs system, which comprises CATs Primary (A0 to A1) for young learners aged 4 to 10, CATs Core (A1 to B1) for school students and young adults, and CATs Higher (B2 to C1) for students in higher education and workplace professionals.

The CATs system uses frequent formative testing of small, achievable, CEFR-aligned learning 'steps', recognising the diversity of learner proficiency, pace and learning style in classrooms. While functional as a standalone system, CATs is designed to supplement conventional classroom instruction, supporting students requiring additional support while stretching those ready for greater challenge.

The pilot involved 2,773 learners across five higher and further education institutions enrolled in CEFR-aligned programmes at levels A1 to B1. Learner engagement in the pilot was strong with a 63% completion rate. Learners demonstrated measurable gains during their use of CATs averaging 48.1 points on the CATs CEFR-aligned scale over 12 weeks, equivalent to almost one-third of a CEFR level. Learner satisfaction was high: 78% wished to continue using CATs and 81% reported perceived improvement in English proficiency. Although gains cannot be attributed exclusively to CATs, the combination of sustained engagement, measured improvement and positive learner perceptions provides encouraging evidence for the approach.

Suggested citation

Milanovic, M., & Jones, C. (2026). *Mobile-based English language learning through bite-sized formative CEFR-aligned assessment*. LANGUAGECERT.

Introduction

The challenge: Disappointing outcomes in formal language education

English language education attracts substantial investment worldwide, yet completion and success rates remain persistently low. The European Survey on Language Competences (ESLC)[1], which studied 50,000 participants from 15 European education systems, illustrates the scale of the problem. Even in Sweden, with its strong English language education tradition, only 57% of students reached CEFR level B2 by age 16 (European Commission, 2012). Across participating countries, outcomes ranged from 2% to 57%, pointing to systemic problems in structure and delivery of English language teaching.

Low success rates are evident across educational contexts, including secondary schools, universities, adult education programmes, and commercial language schools. Many learners give up before achieving their goals, while those who persist often fall short of proficiency targets despite months, or even years of instruction. When designing the CATs system, we therefore asked a fundamental question: how can success rates in CEFR-aligned language education be improved?

Language learning is inherently challenging and time-intensive. Estimates suggest that progression through a single CEFR level typically requires 100 to 200 hours of study, meaning the A1 to B1 journey may require 500 to 600 hours of sustained effort. During this time, learners may study for months without receiving clear evidence of progress and this can create motivational problems. Traditional testing tends to occur infrequently, perhaps once per term and again in a final examination. Limited feedback means learners discover gaps in understanding only after weeks of instruction and this is often too late for remediation. Conventional approaches also tend to separate instruction from testing, with testing positioned as a retrospective measure of learning rather than as an integral part of the learning process itself.

This study reports on a large-scale pilot implementation of the CATs system in operational educational contexts. The study focuses on usability, learner engagement, and indicative learning outcomes in authentic educational settings, rather than on controlled experimental validation. As such, the analyses

presented are descriptive and exploratory in nature, providing a foundation for subsequent, more controlled validation research.

Assessment-oriented learning: the theoretical foundation

The CATs system treats assessment as a primary driver of learning, rather than as a purely evaluative tool. This approach draws on research showing that well-designed testing enhances learning outcomes.

The *testing effect* (also known as retrieval practice) demonstrates that actively retrieving information from memory strengthens learning more effectively than additional study alone (Roediger & Karpicke, 2006). In language learning, retrieving vocabulary, grammatical structures, or communicative strategies through test tasks can support stronger memory and more automatic access than passive review. CATs embeds frequent testing throughout the learning process, transforming every evaluation into a learning event.

Retrieval practice is most effective when it is challenging yet achievable. In other words, learners must work to recall information but ultimately succeed. This principle of "desirable difficulty" (Bjork & Bjork, 2011) shapes how CATs targets appropriately challenging levels. If tasks are too easy, they provide little benefit because recall is automatic. If they are too difficult, they risk frustrating learners without successful retrieval.

Formative assessment research further supports this approach. Black & Wiliam's (1998) meta-analysis of more than 250 studies demonstrated that assessment designed to provide feedback for improvement, rather than grades alone, produces significant learning gains, with effect sizes from 0.4 to 0.7, representing improvement from the 50th to the 65th to 75th percentile. Immediate, specific feedback helps learners identify misunderstandings, correct errors and adjust strategies. Frequent, low-stakes testing informs ongoing learning and can lead to improved performance in high-stakes exams that document achievement retrospectively. CATs operationalizes these principles by providing multiple formative tests at each sub-level.

Mastery learning frameworks, developed through Bloom's (1968) research and later refined by Guskey (2010), show that when instruction is followed by formative testing and opportunities for additional corrective practice, nearly all learners can achieve high levels of mastery. Differences in outcomes reflect differences in time and opportunity to learn rather than innate ability. Language

learning exemplifies this. We all learn at least one language, proving capacity exists. By breaking learning into small units, testing mastery of each, and providing support where needed we can improve success significantly. CATs applies these principles by subdividing CEFR levels into smaller units, each requiring demonstrated mastery before progression. Learners must complete certification tests at each sub-level with more than 70% mastery before advancing, so progression reflects demonstrated competence rather than time spent.

Testing, when frequent, formative, and aligned with clear objectives, drives learning rather than merely measuring it. CATs combines these principles in a mobile-delivered system that makes frequent, high-quality testing practical and scalable.

The CATs innovation: small chunks, frequent testing, demonstrated mastery

What distinguishes CATs is not mobile delivery alone, as many systems offer that. The innovation lies in how CEFR-aligned learning is structured and validated.

CATs Core divides the three CEFR levels from A1 to B1 into nine manageable sub-levels (A1.1 to B1.3). Rather than working toward a single distant goal such as A2 (100 to 150 hours), learners complete three smaller, clearly defined steps: A2.1, A2.2, and A2.3, each representing approximately 40 to 50 hours of learning. This structure transforms overwhelming goals into achievable milestones, enabling learners to see tangible progress within weeks rather than months or years. Regular certification at each sub-level provides frequent validation and sustains motivation, particularly during the foundational A1 to B1 stages where, as the ESLC shows, attrition is highest.

More fundamentally, CATs positions testing as the central organising principle of learning, rather than an occasional interruption to instruction. Each sub-level includes regular formative practice after every two instructional units for feedback on mastery, two complete practice tests simulating certification conditions, and certification tests validating achievement before advancement. Multiple evaluations per sub-level across 40 to 50 hours mean learners engage with testing as continuous learning, not occasional obligation.

The experience differs from traditional instruction-then-test models. Conventional approaches position tests as summative evaluations determining

whether learning occurred. CATs embeds testing throughout, continuously informing and reinforcing learning. When practice reveals struggles with the past tense or listening comprehension of numbers, learners can immediately focus additional practice. Continuous feedback helps learners identify gaps before they compound. Every test measures current competence while strengthening it through measurement.

Mobile delivery makes this test-rich approach practical for supporting classroom instruction. Automated assessment of all four language skills (listening, reading, writing and speaking) provides immediate feedback that would be impractical with human raters. Learners get immediate results, enabling multiple cycles of testing, feedback, and practice in a single session.

Research objectives: Validating the approach

This pilot study evaluates whether CATs pedagogy produces the engagement and learning outcomes predicted by its theoretical foundations. Two primary objectives guided this research.

The first objective addressed practical viability: whether frequent testing of small CEFR-aligned language points produces measurable proficiency gains while maintaining learner engagement? Does the approach work in practice? Will learners accept frequent testing, or will they find it burdensome despite theoretical benefits? Does it drive outcomes and sustain motivation over meaningful timeframes? We examined whether learners completing sub-levels showed proficiency gains, found the test-rich approach motivating rather than stressful, and whether the granular structure reduced dropout relative to typical online learning.

Second, the study aimed to establish a foundation for future validation research examining whether proficiency gains demonstrated within CATs transfer to external measures, particularly international exams referencing CEFR frameworks.

Three research questions guided the analysis:

1. How do learners react to frequent testing (e.g. motivation, perceived usefulness and stress), and how manageable is the system for coordinators, as evidenced through questionnaire data and focus group feedback?

2. Does frequent formative testing of small CEFR-aligned language points produce measurable proficiency gains, as evidenced through system-generated assessment data?

3. Can a test-rich mobile approach work be implemented effectively across diverse institutional contexts, as evidenced through system usage data and coordinator feedback?

Success on these dimensions – learner experience, learning outcomes and institutional feasibility – would suggest this approach has the potential to address persistent challenges in CEFR-aligned language education, with implications for instructional design.

CATs System: design and implementation

Implementing learning theory through system design

CATs operationalizes three complementary learning frameworks: retrieval practice, formative assessment, and mastery learning.

A fundamental principle of the CATs system is that learners completing frequent CATs tests engage in retrieval practice that strengthens competencies better than reviewing grammar explanations or reading examples alone. This aims to leverage the testing effect, where active recall creates stronger memory traces than passive study. The "desirable difficulty" principle (Bjork & Bjork, 2011) guides task design to target appropriately challenging levels. When tasks are too easy, they provide little benefit because recall is automatic. When they are too difficult, they can be a cause of frustration.

Black and Wiliam's (1998) work on formative assessment explains why CATs design works. Their meta-analysis of more than 250 studies found interventions emphasising formative testing improved performance from the 50th to the 65th to the 75th percentile. Key characteristics align with how CATs operates – tests occur frequently, not occasionally; results arrive immediately, not after delays; learners can act on feedback through additional practice.

Bloom's mastery learning principle operates through the sub-level structure. Learners must complete certification tests at each sub-level, demonstrating more than 70% mastery before advancing. Progression depends on demonstrated competence, not time spent. A learner's B1.2 certificate indicates

actual B1.2 competence, unlike traditional course completion certificates that may or may not reflect proficiency. So, progressively through any given level, the aim is to firm up the foundation so that addressing the next rung on the ladder comes from a position of competence.

All three frameworks (i.e., retrieval practice, formative assessment, and mastery learning) point to the same conclusion: testing, when frequent, formative, appropriately challenging, and connected to clear mastery criteria, becomes a powerful learning mechanism. CATs builds an entire system around this idea, making testing the continuous backbone of learning rather than an occasional event.

System architecture

The CATs system comprises four integrated components that work together to deliver learning at scale.

Component 1 – Granular CEFR progression

CATs subdivides levels A1 to B1 into nine sub-levels (A1.1, A1.2, A1.3, A2.1, A2.2, A2.3, B1.1, B1.2, B1.3), addressing motivation and persistence challenges inherent in lengthy language programmes. Full CEFR levels represent daunting goals requiring 100 to 200 hours over months or years. Sub-levels transform these into achievable milestones reachable in 6 to 8 weeks with each corresponding to roughly 40 to 50 hours of content and concluding with certification validating mastery. This structure provides regular validation through certificates documenting specific achievements, rather than extended periods without tangible progress markers. The psychological impact is significant: learners see concrete advancement every few weeks rather than waiting months or years.

Component 2 – Preparation materials

Each sub-level provides 30 topic-based units addressing real-world communication situations. Unlike traditional approaches where content is simply delivered for absorption, CATs materials prepare learners explicitly for assessment. After every two units (approximately 3 to 4 hours of learning), practice activities test recently covered material, providing immediate feedback and identifying areas requiring focus. Supplementary modules (Vocabulary Extra, Grammar Extra, Video Extra) provide targeted practice for specific competencies evaluated in certification tests. These resources align with test requirements,

enabling learners to focus on the skills certification will evaluate. Two complete practice tests per level familiarise learners with formats, reduce anxiety and allow self-evaluation of readiness.

Component 3 – Multi-layered testing framework

CATs embeds testing throughout the learning experience at multiple layers, each serving distinct purposes:

Diagnostic placement occurs at programme entry, using adaptive algorithms to determine each learner's proficiency within 20 to 30 minutes, ensuring learners begin at appropriate levels to maximize efficiency and maintain engagement. Placement that is too high can cause frustration; placement that is too low wastes time on mastered material.

Formative practice occurs after every two instructional units, providing regular feedback about mastery. These low-stakes assessment opportunities strengthen learning through retrieval practice, identify gaps, validate understanding before learners build further on that foundation, and maintain progress awareness. The regular rhythm (every 3 to 4 hours) prevents accumulating misunderstandings.

Practice tests (two per sub-level) simulate certification conditions across all four skills under realistic time constraints. These serve preparatory and evaluative functions: familiarising learners with test formats and timings, identifying areas requiring additional preparation, and building confidence through successful performance.

Certification tests validate mastery at completion of each sub-level, requiring 70%+ performance across all four skills before advancement. These evaluations gate progression, ensuring it reflects demonstrated competence. A learner achieving B1.3 certification has achieved B1.3 competence as CEFR defines it.

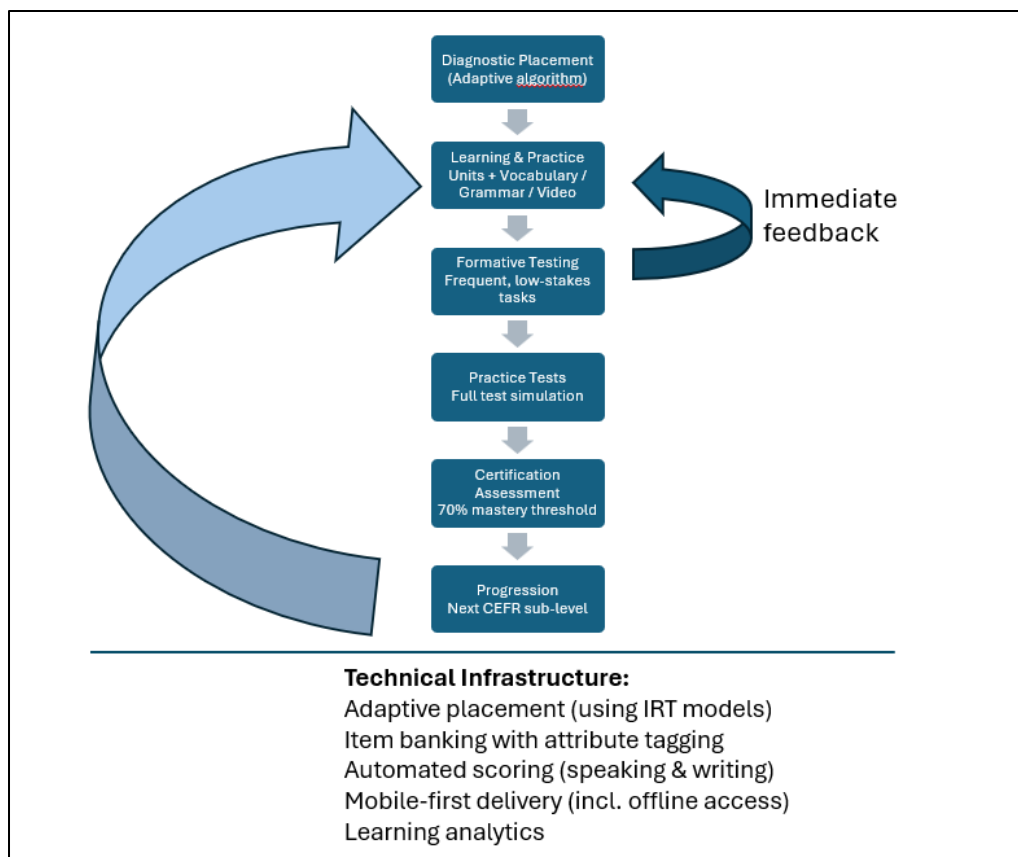
Component 4 – Technical infrastructure

CATs uses automated scoring across all four skills, including speaking, via speech recognition and writing via algorithmic evaluation. This provides immediate, detailed feedback, making frequent, high-quality testing practical and affordable. Learners receive results immediately rather than waiting days, enabling multiple cycles of testing, feedback, and practice within single sessions.

Sophisticated item banking systems with detailed attribute tagging ensure tests target specific competencies at appropriate difficulty levels. Adaptive algorithms adjust difficulty in response to performance patterns, maintaining optimal challenge throughout progression.

Learning analytics capture detailed information about performance patterns, enabling both individual feedback and system-level refinement. Learners receive specific guidance about strengths and development areas. Coordinators can identify learners struggling with particular competencies and provide targeted support. System developers can identify items or materials requiring refinement. This data infrastructure supports continuous improvement in both outcomes and system quality. Figure 1 provides an overview of the CATs system architecture and learning cycle, illustrating how adaptive placement, structured learning and multi-layered assessment are integrated within a continuous feedback loop. In this model, assessment functions not as a terminal event but as an ongoing mechanism of learning, with system infrastructure enabling this process at scale.

Figure 1. Overview of the CATs system architecture and learning cycle



CATs pilot study methodology

Research design

The pilot study employed a mixed-methods design combining quantitative analysis of learner performance data with qualitative evidence from questionnaires and focus groups. The research design enabled evaluation of both learning outcomes and implementation experiences.

Participants and context

A total of 2,773 learners from five higher and further education institutions participated in the pilot study. Institutions included two community colleges and three polytechnic institutions, with cohort sizes ranging from approximately 100 to more than 1,100 learners. This diverse institutional representation enabled analysis across differing organisational contexts and baseline proficiency profiles.

Representatives from each institution attended remote training sessions covering system administration, learner registration and progress monitoring. Two central pilot coordinators, based within participating institutions, facilitated regular collaboration meetings and provided ongoing implementation support.

The CATs intervention

The pilot deployed CATs Core (A1.1 to B1.3) as described in Section 2, including preparation materials, practice activities, practice tests and certificated assessments. Each sub-level provided 40 to 50 hours of structured content organized into 30 topic-based units with access to supplementary modules (Vocabulary Extra, Grammar Extra, Video Extra).

The system's technical infrastructure included adaptive placement algorithms using item response theory (IRT), automated speech recognition for speaking assessment, and algorithmic scoring for writing assessment.

Implementation process

The pilot followed four phases designed to minimize institutional burden while maintaining consistent procedures across sites and ensuring data quality:

Phase 1: Institutional setup and training

Representatives from participating institutions attended remote training sessions covering system administration and learner support procedures. Each institution designated coordinators responsible for code distribution and progress monitoring.

Phase 2: Learner registration and placement

Learners registered using institutional codes, completed demographic questionnaires, and undertook adaptive placement testing. Placement results determined appropriate starting levels.

Phase 3: Learning engagement

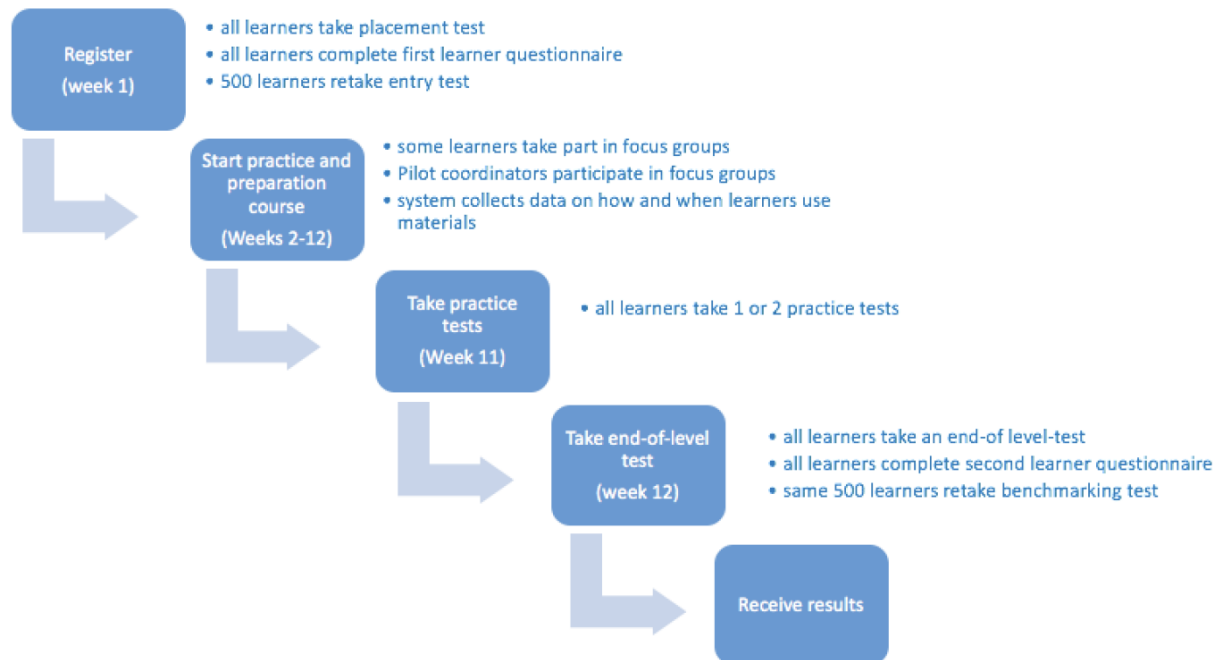
Learners accessed preparation and practice materials via a mobile phone or tablet over a 12-week period. Web-based delivery was not available during this pilot. The platform recorded usage data including time spent, completion rates, and performance on practice activities.

Phase 4: Assessment and certification

Learners who completed preparation and practice tests undertook final certification assessments. All assessments were automatically scored with results linked to learner profiles for analysis.

Figure 2 provides an overview of the pilot implementation process, illustrating how learning, assessment and data collection were integrated across the 12-week study period to support both learner progression and evaluation of the system.

Figure 2. Overview of the CATs pilot implementation process



Data collection

Four principal sources of data were used:

Learner questionnaires

Learners completed two online questionnaires, collecting demographic information, learning background, and evaluation of system effectiveness and satisfaction.

Learners completed the first learner questionnaire at the point of registration. This short questionnaire asked for some basic personal details (e.g. gender, date of birth, department and subject, national examination result for English) and information about learners' English-language context (e.g. reasons for learning English, years learning English, time spent learning English per week, access to English-language materials and media). The first questionnaire asked students to rate on a scale of 1 to 5 how difficult they find different aspects of English language learning. This question was later repeated in the second questionnaire.

The second learner questionnaire was completed at the end of the pilot, after the end-of-level final test. This longer questionnaire included further questions about the learner (e.g. level of schooling, English proficiency of parents) and questions about the user experience learning with CATs (e.g. how effective, relevant and engaging learners found different aspects of the materials and the system).

Focus groups

Separate focus group sessions were conducted with learners and institutional coordinators to gather qualitative feedback on user experience, system effectiveness, and implementation challenges.

Learner focus groups took place eight weeks into the pilot. These explored the extent to which CATs supported English-language learning and the specific ways in which learners think this happens. Learner focus groups collected qualitative feedback on the user experience, the functionality of the system, the usability of the materials, the language activities learners find most helpful, and individual learning styles.

Pilot coordinator focus groups were designed to explore the experience of managing CATs from the institution administrator's point of view and collect qualitative feedback on running and delivering CATs. Two focus groups took place with pilot coordinators from each institution. The first took place eight weeks into the pilot and the second at the end of the pilot process. The focus groups were facilitated by the CATs Team and conducted remotely.

System-generated data

The CATs platform automatically captures comprehensive data on learner activity, including time spent, completion rates, assessment scores, and detailed interaction patterns.

During the pilot the following information was captured and stored for all learning and assessment materials:

- User System Fields (Unique CATs User Identifier, System Identifier, Device Identifier)
- Question Data (Task ID, Question ID, Answer given, Score, Maximum Possible Score)
- Task Data (Task ID, Timestamp, Score, Maximum Possible Score, Number of times this task has been opened, Open duration)

- Task Progress (Unit ID, Task ID, Maximum Score, Latest Score, Latest Score Date, Highest Score, Highest Score Date)
- Unit Progress (Unit ID, Unit Score, Unit Maximum Score, Last Attempt Date).

Assessment scores

Pre- and post-intervention assessment scores provided quantitative measures of learning effectiveness, with all assessments calibrated to CEFR levels using IRT scaling methodology.

The CATs platform captures extensive granular data on learner interactions and performance. While aggregated metrics were available for the present analysis, more detailed exploitation of item-level and interaction-level data was beyond the scope of this study and represents an important direction for future research.

Data analysis

The data analysed in this study were generated through system use and learner feedback mechanisms within an operational pilot. As such, they were not originally collected within a controlled experimental design. The analysis therefore focuses on descriptive patterns in engagement, performance and user experience, rather than on hypothesis testing or causal inference.

Quantitative analysis focused on learning gain measures, with proficiency improvement calculated as the difference between adaptive placement test scores and end-of-level assessment scores on the 0–900 CATs scale. The CATs scale is constructed using Rasch measurement methodology, with scale values representing positions on a unidimensional proficiency continuum. A difference of 50 points on this scale corresponds to progression through one CATs sub-level, which represents one-third of a full CEFR level.

Qualitative data from focus groups and questionnaires were analysed thematically to identify key factors influencing implementation success and learner satisfaction.

Results

Participation and engagement

1,738 of 2,773 registered learners (62.7%) completed end-of-level assessments, comparing favourably with typical online learning completion rates. Learners completed an average of 31% of available materials overall (48% excluding non-engagers). Critically, learners completing the first 10% of materials showed significantly higher continuation rates, indicating the importance of initial engagement for sustained preparation effort.

Baseline proficiency assessment

Placement test results revealed proficiency distributions consistent with the target population for CEFR levels A1 to B1. Learners from polytechnic institutions typically demonstrated proficiency around CEFR levels A2/B1, while those from community colleges showed slightly lower baseline levels. The placement testing successfully differentiated learners across the target proficiency range, with appropriate score distributions across the nine CATs sub-levels. Standard deviations indicated substantial variation in ability within institutions, supporting the value of individualized placement rather than institution-wide approaches.

Learning outcomes: Proficiency improvements

The primary outcome measure, change in proficiency between placement and end-of-level assessments, demonstrated substantial learning gains.

Overall learning gains

Mean gain (CATs scale)	48.1
Equivalent CEFR progress	~0.33 level

Learners achieved an average improvement of 48.1 points on the CATs measurement scale over 12 weeks. Since the CATs scale is calibrated such that 50 points represents one CATs sub-level (equivalent to one-third of a CEFR level), this represents progress of approximately 0.96 CATs levels, or nearly one-third of a full CEFR level. While gain scores provide a useful indicative measure of progress within the CATs scale, they should be interpreted with caution. More robust approaches to measuring learning impact, including controlled comparisons and external validation against independent assessments, are required to draw definitive conclusions about effectiveness.

Distribution of improvements

Analysis of learner distribution across proficiency levels showed clear evidence of upward progression. Between placement and end-of-level testing, substantially fewer learners remained at A2 level, with greater proportions progressing to B1 or B2 levels.

Figure 3 illustrates the distribution of learners across CEFR-aligned sub-levels at placement and end-of-level assessment. The observed shift towards higher levels is consistent with the gain scores reported, although, as noted, these changes cannot be attributed solely to CATs.

Figure 3. Distribution of learner proficiency levels at placement and end-of-level assessment

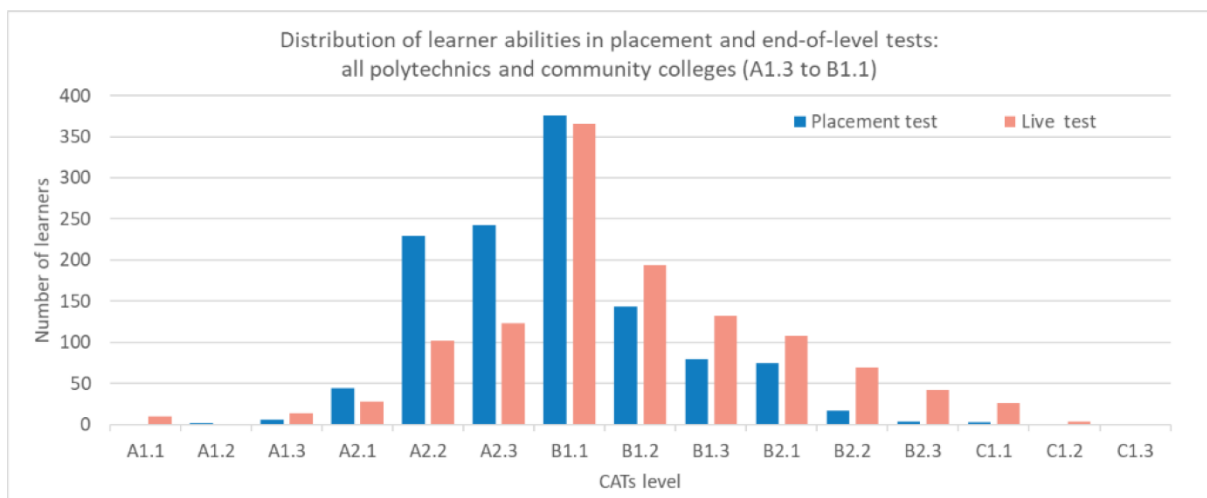
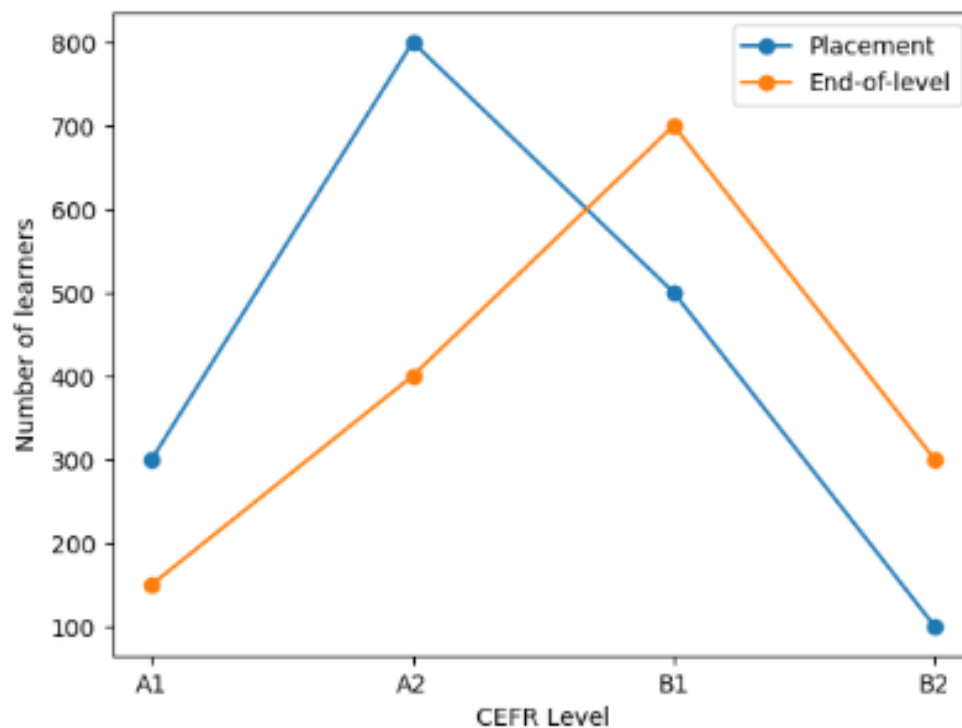


Figure 4 presents an aggregated view of learner distribution across CEFR levels at placement and end-of-level assessment. The observed shift towards higher levels is consistent with the gain scores reported.

Figure 4. Distribution of learner proficiency: placement vs end-of-level



Institutional variation

The greatest gains were observed among learners at community colleges, with one institution demonstrating an average improvement of 83 points, equivalent to nearly two-thirds of a CEFR level. This pattern suggests that mobile-accessible preparation may be particularly beneficial for learners who face greater barriers in traditional preparation contexts.

Assessment performance quality

End-of-level test scores reinforced the proficiency gains observed through the measurement scale. Average raw scores ranged from 38 and 40 out of 50 possible points, with most learners achieving scores above 70%. These results exceeded typical mastery thresholds and were observed consistently across all tested levels, indicating that learners developed genuine competence within the framework assessed by CATs.

Learner experience and satisfaction

Despite relatively low response rates to end-of-pilot questionnaires, the available data and supporting qualitative feedback provide useful insights into learner perceptions of the system. Available satisfaction data indicated highly positive learner experiences: 78% of respondents expressed desire to continue using the CATs system, 76% reported enjoying use of the application and 81% believed their English proficiency had improved through system use.

Focus group discussions revealed that learners particularly valued immediate feedback on practice activities, flexible access enabling study at convenient times and locations, clear progress indicators showing advancement toward proficiency goals and variety in learning activities maintaining engagement over extended periods. These satisfaction indicators suggest that preparation approaches like CATs can sustain the extended engagement necessary for effective language development.

Taken together, these findings suggest that learners particularly value immediacy of feedback, flexibility of access and clear indicators of progress, all of which are central to the CATs design.

Institutional implementation feasibility

Focus groups with institutional coordinators provided insights critical for understanding scalability potential:

Technical implementation

Coordinators reported successful system implementation across diverse institutional contexts with minimal technical difficulties. The registration process functioned smoothly even with large student cohorts, and internet bandwidth concerns proved largely unfounded. This technical reliability supports institutional confidence in the system.

Administrative efficiency

After initial training, coordinators found they could manage the system with minimal ongoing support from technical staff. This finding has significant implications for scalability, as it suggests that large-scale implementation would not require extensive specialised infrastructure.

Learner support requirements

While most learners engaged successfully with materials, coordinators identified that students with initially low motivation required additional encouragement. They suggested that automated progress notifications and social media integration could enhance engagement.

Monitoring capabilities

Coordinators found the learner monitoring system effective for tracking both individual and group progress. The ability to identify learners requiring additional support proved valuable for pastoral care.

System reliability and technical performance

The CATs platform demonstrated high reliability throughout the pilot period, with no major technical failures or system crashes reported. This stability proved crucial for maintaining learner engagement and institutional confidence. The mobile-first design proved successful in accommodating diverse learning contexts, with successful engagement across various device types and demonstrated feasibility of offline content access in contexts with limited connectivity.

Discussion

CATs effectiveness for English language development

The average improvement of 48.1 points represents meaningful progress approaching one-third of a CEFR level within 12 weeks, comparing favourably with published research on similar interventions. This suggests that well-designed mobile systems can produce measurable outcomes within reasonable periods. Gains occurred consistently across institutional contexts and proficiency levels, with particularly strong results among community college learners, suggesting mobile-based learning may especially benefit learners facing traditional access barriers.

Theoretical implications for CEFR-aligned examination preparation

CATs and major CEFR-aligned examinations share construct alignment through common CEFR descriptors and emphasis on communicative competence. This theoretical foundation suggests competencies developed through CATs should transfer to examination performance. However, this requires empirical validation examining:

Assessment format correspondence

While CATs and major examinations share CEFR alignment, specific task types and response formats may differ. Research should examine whether competencies developed through CATs automated assessment tasks transfer to different assessment formats.

Proficiency range considerations

The current pilot examined effectiveness within A1 and B1 levels. Transfer relationships at higher proficiency levels (B2 and C2) remain empirically unexamined and may differ given the increased complexity of linguistic and communicative demands.

Preparation duration adequacy

While 12 weeks proved sufficient for measurable gains on CATs assessments, optimal preparation periods for specific examination targets require investigation, particularly considering that progression through full CEFR levels typically requires substantially longer engagement.

Mobile delivery advantages for language learning access

The successful mobile-first implementation addresses several key challenges in language learning access. The elimination of geographic and scheduling barriers through mobile access significantly expands potential learner reach, particularly benefiting learners in remote areas or those with complex work/study schedules.

The individualized pacing enabled by mobile delivery contrasts favourably with fixed-schedule classroom instruction. Learners could adjust their learning intensity based on their circumstances, enabling both intensive short-term engagement and extended skill development approaches.

The high learner satisfaction rates suggest that mobile-based learning can sustain the extended engagement necessary for meaningful language development. The immediate feedback and progress tracking features appeared particularly valuable for maintaining motivation over extended periods.

Scalability for institutional implementation

The minimal institutional implementation burden reported by coordinators indicates strong scalability potential. The ability to manage large cohorts with minimal specialized staff suggests that CATs-based learning could be efficiently delivered across diverse institutional contexts.

The system reliability and technical performance demonstrated throughout the pilot support confidence in large-scale delivery. The absence of major technical issues and successful accommodation of varying connectivity contexts indicate robust platform capabilities.

The positive coordinator feedback regarding monitoring and support capabilities suggests that institutional implementation of CATs provides adequate oversight and learner support mechanisms for responsible program delivery.

Study limitations and future research directions

Several limitations affect the interpretation and generalizability of these findings:

Absence of external validation

A key limitation of this study is the absence of external validation measures. All learning gains were assessed using the CATs system itself and, while these assessments are CEFR-aligned and psychometrically informed, the findings cannot yet be directly linked to performance on independent CEFR-aligned examinations. Establishing such relationships represents a critical next step for evaluating the broader validity of the approach.

Limited proficiency range

The pilot focused on learners within A1 to B1 CEFR levels. Conclusions about CATs effectiveness for higher proficiency levels cannot be drawn from these data. The linguistic and communicative demands at B2 to C2 levels differ substantially, and effectiveness at these levels requires separate investigation.

Timeframe constraints

The 12-week study period, while sufficient to demonstrate learning gains, does not address longer-term retention or sustained proficiency development. Longitudinal research examining retention and continued progression would provide a more complete understanding of learning trajectories.

Completion bias

The 62.7% completion rate, while reasonable for online learning contexts, means that findings represent learners who sustained engagement with the system. The characteristics and outcomes of non-completing learners remain less well understood, which may limit the generalisability of the results.

Cultural and linguistic context

All participants were drawn from specific institutional and geographic contexts. Effectiveness may vary across different cultural settings, first language backgrounds and educational traditions, and further research is needed to explore these dimensions.

Concurrent learning effects

Learners continued their regular classroom studies while engaging in the CATs pilot. As a result, it is not possible to isolate the specific contribution of CATs to the observed improvements and causal inferences regarding its impact cannot be drawn.

Data scope and analytical depth

While the CATs platform captures extensive granular data on learner interaction and performance, the present study draws primarily on aggregated metrics generated during the pilot. More detailed analyses of item-level performance, learning behaviours and interaction patterns would enable a deeper understanding of how specific features of the system contribute to learning.

Future research directions

Taken together, these limitations indicate that the current findings should be interpreted as indicative rather than definitive. This study establishes an operational and empirical baseline for future research, which should focus on controlled designs, external validation against independent assessments and deeper analysis of learner interaction data to more fully evaluate the effectiveness of the CATs approach.

Conclusions

This study provides indicative evidence that the CATs mobile-based learning system effectively supports English language development within the A1 to B1 CEFR proficiency range. Learners achieved substantial proficiency improvements over a 12-week period, with learning gains averaging 48.1 points on a CEFR-aligned scale, representing approximately one-third of a CEFR level. High learner satisfaction combined with minimal institutional implementation burden suggests strong potential for scalable delivery.

Implications for CEFR-aligned examination contexts

The shared CEFR framework foundation between CATs and major international English language examinations, including those offered by LANGUAGECERT creates theoretical conditions for transfer of competencies developed through CATs to examination performance. The emphasis on communicative competence within authentic contexts in both systems supports this theoretical alignment.

However, the critical question of whether proficiency gains demonstrated through CATs assessments transfer to performance on external CEFR-aligned examinations remains empirically unresolved. While theoretical construct alignment and successful learning outcomes create favourable conditions for such transfer, definitive conclusions require direct validation through examination performance studies.

Contributions and future directions

This study contributes to mobile-assisted language learning literature by providing empirical evidence from a large-scale pilot implementation demonstrating both learning effectiveness and implementation feasibility. The findings establish CATs as a viable mobile-based learning system with demonstrated impact on language proficiency development.

For the broader field of English language learning and assessment, these results demonstrate that mobile-based approaches can produce meaningful learning outcomes while addressing practical barriers of access, cost, and scheduling flexibility. The combination of CEFR alignment, comprehensive skill development, and flexible delivery creates a learning framework that addresses both pedagogical effectiveness and practical implementation challenges.

This study establishes an operational and empirical baseline for future validation work. The next phase of research will focus on linking CATs engagement and performance to outcomes on external CEFR-aligned assessments, enabling a more robust evaluation of its effectiveness as a preparation tool. Longitudinal studies tracking learners from CATs participation through to examination outcomes would provide more definitive evidence of its value as examination preparation and would complete the empirical foundation for mobile-based approaches.

Until such validation studies are completed, claims regarding CATs effectiveness as preparation for specific examinations must remain qualified. The current evidence supports CATs as an effective language learning system with strong theoretical alignment to examination constructs, establishing a foundation that warrants, and indeed requires, further investigation through direct examination performance research.

References

- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.
<https://doi.org/10.1080/0969595980050102>
- Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.
- Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, 1(2), 1–12.
- European Commission. (2012). *First European survey on language competences: Final report*. Publications Office of the European Union.
- Guskey, T. R. (2010). Lessons of mastery learning. *Educational Leadership*, 68(2), 52–57.
- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255.
<https://doi.org/10.1111/j.1467-9280.2006.01693.x>



by PeopleCert

MILLIONS OF EXAMS DELIVERED WORLDWIDE

LANGUAGECERT is an Awarding Organisation recognised by Ofqual. It spearheads innovations in language assessment and certification, providing high-quality services to the global learners' community. It is a UK-based member of PeopleCert Group, a global leader in the certification industry, that delivers millions of exams in over 200 countries.

Learn more about LANGUAGECERT exams at:
www.languagecert.org

LANGUAGECERT is a business name of PeopleCert
Qualifications Ltd, UK company number 09620926

✉ info@languagecert.org

f [/LanguageCert.org](https://www.facebook.com/LanguageCert.org)