African Educational Research Journal
Vol. 14(3), pp. 707-716, September 2026
ISSN: 2354-2160
Full Length Research Paper

Automated online learning in pre-service teacher education: A one-group study of recognition-based learning outcomes and estimated workload substitution

Theerapat Sinthudech, Paweena Khansila and Anucha Pimsak*

Faculty of Education and Educational Innovation, Kalasin University, 13 Moo 14, Song Plueai Sub-district, Na Mon District, Kalasin 46230, Thailand.

Accepted 2 September, 2026

ABSTRACT

Faculties of education are increasingly required to develop competences for which they have no staffing capacity, and automated online provision is the usual response. This study asks not whether such provision works but how much staff workload it can be estimated to absorb, and whether that absorption is even across a language competence. The course was designed around self-directed learning, with staff facilitating rather than being removed. An online micro-credential in Thai professional communication comprising 22 lessons was developed through educational design research and delivered to pre-service teachers across three refinement cycles without live instruction, with all assessment machine scored. The analysed sample comprises 128 pre-service teachers, all of whom completed all 22 lessons within the allocated time. Against the faculty workload framework, conventional delivery in four sections would have required 280 contact hours and the marking of 2,944 papers, equivalent to 147 to 245 hours, a cost the faculty could not meet and the reason the competence had gone undeveloped. Learning gains on recognition-based tests were substantial, with post-instruction scores of 49.37 exceeding pre-instruction scores of 31.66 out of 80, $t(127) = 30.03$, $p < .001$, $d = 2.25$, 95% CI [16.54, 18.87]. Lessons targeting productive skills showed a lower mean first-attempt pass rate than receptive lessons, 82.4% against 86.8%, but the difference was not significant, $U = 64.0$, $p = .605$. We report the recognition-production boundary as a hypothesis warranting direct test, and distinguish estimated workload substitution from realised saving throughout.

Keywords: Teaching capacity, automated assessment, online course delivery, recognition-based measurement, teacher education, educational design research.

*Corresponding author. E-mail: anucha.pi@ksu.ac.th. Tel: +66 43 602 037.

INTRODUCTION

Faculties of education across many systems face a structural mismatch. Professional standards specify competences that graduates must demonstrate, while the staffing establishment and the credit structure through which those competences would conventionally be developed remain fixed. The response most often proposed is online provision, and the question most often asked of it is whether it produces learning. Meta-analytic evidence indicates that it generally does (Morina et al., 2025), with a mean teacher-level effect of 0.71 across 102 pre-test/post-test studies of online teacher professional development, and with control-group designs yielding significantly smaller effects than within-subject designs.

We suggest that this framing obscures the decision institutions actually face. No faculty proposes to dispense with its academic staff. What faculties do is decide which components of provision can be carried by a system so that staff time is available for components that cannot. The useful question is therefore not whether automated provision replaces teachers but how much workload it can be estimated to absorb, and whether that absorption is even across what a course teaches. We term the first quantity estimated workload substitution and distinguish it throughout from realised saving, which this study did not observe.

The conditions producing this question are widespread.

Teacher shortage and constrained staffing are reported across systems at very different levels of resource (UNESCO, 2023), and regulators have generally specified outcomes rather than funding the capacity to reach them. Institutions therefore adopt automated provision under conditions where the alternative is not a better-resourced conventional course but no provision at all. The relevant comparison is accordingly not between online and face-to-face teaching of equal quality but between automated provision and its absence.

Why absorption might be uneven

There is reason to expect that automation will not serve all components of a language course equally, and the expectation has both an assessment basis and a cognitive one. On the assessment side, machine scoring requires items with a single defensible correct answer, a condition that receptive skills satisfy readily and productive skills do not. Writing is assessed adequately only where a learner produces text that a reader evaluates against criteria of appropriateness, organisation and register (Weigle, 2002), and speaking imposes the analogous requirement of a listener. Automated writing evaluation has advanced considerably since early systems (Dikli, 2006; Warschauer and Grimes, 2008), and contemporary engines achieve high agreement with human raters on holistic scores (Attali, 2016). Even so, agreement on a holistic score is not the same as evaluation of communicative appropriateness in a professional situation, and systematic reviews of automated feedback report that gains concentrate in surface features rather than in content and organisation (Fleckenstein et al., 2023).

On the cognitive side, the distinction between recognising an appropriate formulation and producing one corresponds to the distinction between declarative and procedural knowledge (Anderson, 1982; DeKeyser, 2015). Declarative knowledge of a rule can be acquired through exposure and tested by selection; procedural skill in applying it develops through repeated production under conditions approximating use, and its assessment requires production. Multiple-choice items address the first and are silent on the second. Messick (1995) treated the narrowing of a construct to what an instrument can capture as construct under-representation and identified it as a threat to validity equal in standing to the introduction of irrelevant variance. Applied here, the expectation is that a course delivered and assessed automatically will serve listening and reading better than speaking and writing, and that the difference should be observable in indicators independent of the achievement tests.

The diglossic setting

The setting adds a further consideration. Learners in the participating faculty use the Isan dialect in everyday interaction while professional practice requires Standard Thai, a diglossic rather than second-language situation (Ferguson, 1959). In diglossia, the two varieties are functionally distributed rather than sequentially acquired, so learners possess extensive receptive familiarity with the high variety through education and media while having limited occasion to produce it. This asymmetry between reception and production is a feature of the setting before any instruction occurs, and it means that the receptive-productive distinction examined here is not merely a property of the assessment mode but is also present in the learners' prior competence. Where a diglossic population is taught through a mode of assessment that privileges recognition, the two asymmetries may compound.

Research questions

Three questions guide the article. First, how much staff workload can the provision be estimated to have absorbed, and what learning outcome accompanied it? Second, was the estimated absorption even across receptive and productive components of the competence? Third, what conditions govern the transfer of these findings to other settings?

MATERIALS AND METHODS

Design and setting

The study formed part of a three-phase educational design research project conducted between March 2025 and July 2026 (McKenney and Reeves, 2019). The field trial employed a one-group pre-test post-test pre-experimental design administered across three refinement cycles with the same cohort (Campbell and Stanley, 1963). The absence of a comparison group is a defining constraint on what the study can claim and is treated as such throughout. The study was reviewed and approved by the Human Research Ethics Committee of Kalasin University under certificate of approval number HS-KSU 051/2567, issued on 19 November 2024 following exemption review. Written informed consent was obtained from every participant by a research assistant who had no role in course assessment, and research data were held separately from course grading records.

The provision

The course comprised 22 lessons totalling 70 nominal learning hours, organised by the four skills and sequenced from receptive to productive, and was delivered entirely on the institutional learning management system without timetabled sessions. Lessons targeting listening numbered 5, speaking 8, reading 3 and writing 6. Each

lesson followed a fixed structure of five components, namely a summary of the principles addressed, stated learning outcomes, an exposition, a worked professional scenario, and an end-of-lesson test of 20 multiple-choice items presenting situations drawn from teaching practice. Learners had a single attempt at each lesson test and had to reach the pass threshold before the next lesson became available, following mastery learning conventions. Materials were drawn from documents that teachers encounter in service, including lesson plans, post-teaching reflections, student reports and official correspondence.

The delivery model was a design commitment rather than a technical accommodation. The course was built on the premise that competence of this kind develops through self-directed practice and that the professional standard it addresses is one graduates must continue to meet across a career, which makes the capacity to direct one's own learning part of what the course should develop (Knowles, 1975). Staff accordingly took the role of facilitator: enrolling learners, monitoring progress through the platform record, responding to individual queries, and intervening where the record showed a learner stalled. What the design removes is not the teacher but the requirement that learning proceed at a pace set by the timetable rather than by the learner.

The three refinement cycles

Each cycle comprised pre-instruction measurement, delivery, post-instruction measurement and a learner reflection round. Themes reported by at least a quarter of respondents were brought to the development panel, which decided revisions and documented the reasoning. The procedure was identical in all three cycles. Table 1 sets out what learners reported, what was revised and what followed for the measurement.

Two features of Table 1 bear on the interpretation of later results. First, the revision between cycles 1 and 2 raised item difficulty in response to learner reports that items were too easy, and every measurement quality indicator moved favourably while the score proportion fell. Second, pre-instruction means rose across cycles, by 1.65 marks between cycles 1 and 2 and by 11.11 marks between cycles 2 and 3, the latter interval being the one in which learners received training in operating the platform. Cycle 3 pre-instruction scores therefore no longer function as a baseline, and results from that cycle are reported as cumulative rather than as an estimate of what one course completion achieves.

Table 1. The three refinement cycles: learner-reported problems, revisions made, and measurement outcomes (n = 128).

Cycle	What learners reported	What was revised	Score proportion (%)	SD	Pre-post r	At/above 80% (%)	Declining learners
1	Lessons and materials too long; no guidance on using the platform	Lessons and media shortened	70.0	5.98	0.94	10.9	0
2	Test items too easy; materials not matched to small schools	Item difficulty raised; materials contextualised	61.7	8.24	0.64	4.7	7-17
3	Platform difficult to operate; work lost when a session dropped	All revisions integrated; platform training added before testing	93.5	4.06	0.22	99.2	0-1

Score proportion is the post-instruction mean as a percentage of the 80-mark maximum. SD and r refer to post-instruction scores and to the pre-post correlation. Declining learners are within-subject declines, reported as a range across the four skills. Cycle 3 pre-instruction scores follow platform training and do not function as a baseline.

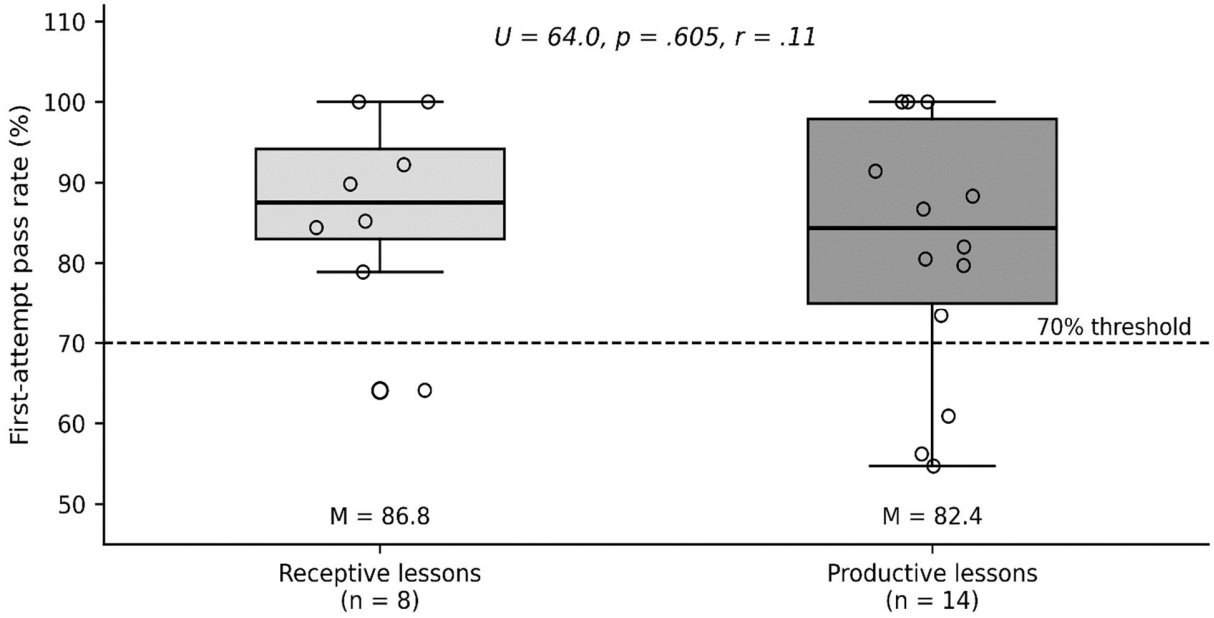

Figure 1. Distribution of first-attempt pass rates for receptive and productive lessons. Each point is one lesson. The difference between groups is not statistically significant.

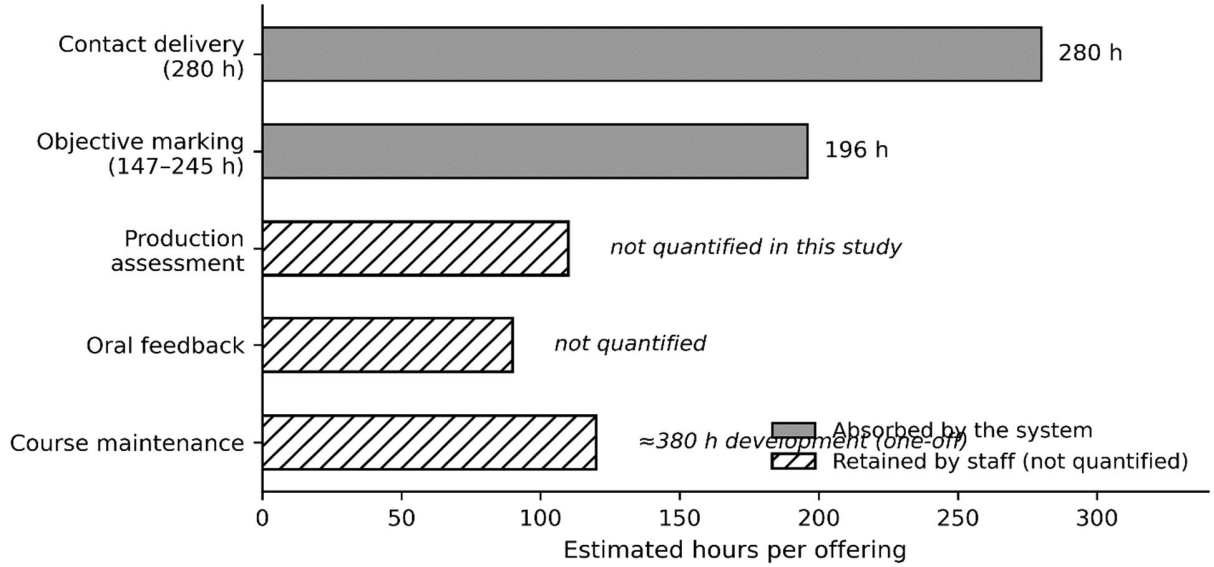

Figure 2. Estimated workload by function, distinguishing what the system absorbed from what remained with staff. Retained functions were not quantified in this study.

Participants and the sequence of trials

The design research process placed two groups of learners in two distinct roles, and the distinction matters for reading the participant numbers. In phase 2.2, a small-scale tryout of version 1 of the prototype was conducted with 168 students enrolled in a general education course, purposively selected because their year of study, class structure and resource conditions resembled those of the target population. The tryout ran for eight weeks between December 2025 and February 2026 and produced significant gains, $t(167) = 35.21$, $p < .001$, $d = 2.88$, together with systemic evaluation ratings averaging 4.117. Its purpose was formative: it generated 12 revisions, including a reduction of the curriculum from 100 to 70 hours, which produced version 2.

In phase 3.1, version 2 was taken into field trial with 128 pre-service teachers drawn from 12 subject majors, the population to which the professional standard applies. This is the group whose results are reported in this article. All 128 participated in all three refinement cycles and all completed all 22 lessons; the platform requires a learner to pass each lesson before the next becomes available and issues the course record only on completion, so partial completion does not arise within this group. Learners entered and finished at different times according to their own schedules, which the self-paced design permits, but none exceeded the time the curriculum allocates.

Because the platform holds records for every learner who has ever been enrolled, its totals span both trials and a third group of self-enrolled users, and therefore require disaggregation. Of the 305 accounts on the system, 7 belonged to experts trialling the platform for evaluation purposes and 2 to the researchers, leaving 293 learner accounts. Of these, 176 had completed all 22 lessons: 128 pre-service teachers from the phase 3.1 field trial and 48 general education students remaining from the phase 2.2 tryout. The reduction from 176 to 128 is a separation of two groups belonging to different phases of the design research process, not attrition within the analysed sample.

The remaining 117 accounts belong to a third group. The platform was open to the institution rather than restricted to the trial, and the course represented the first occasion on which a test of Thai listening, speaking, reading and writing had been made available through an institutional system. Pre-service teachers outside the trial cohort enrolled voluntarily on that account, the professional standard being one on which their progression depends. These learners were not assigned to any trial, were not measured before or after instruction, did not complete the reflection instrument, and are excluded from every analysis reported here. Their enrolment is reported for completeness of the platform record and because voluntary uptake on this scale, from learners under no obligation to participate, indicates that the provision addressed a need the population recognised. No claim is made about their learning, since none was measured.

A post hoc calculation indicates that with n = 128 and alpha of .05, the design had power exceeding .99 to detect a within-subject effect of d = 0.5, so it was not underpowered for the primary comparison. The lesson-level comparison between modalities, with 8 and 14 lessons, had power of approximately .13 to detect a difference of the magnitude observed, and is discussed accordingly.

Estimating absorbed workload

Two quantities were estimated, and both are estimates against a counterfactual rather than observations of savings. The counterfactual is anchored in the workload framework the faculty applies to course design. The provision reported here is a micro-credential and carries no credit of its own, but it was designed against the faculty template for a three-credit course, and that template supplies the basis for the estimate. Under it, a three-credit course carries nine timetabled hours per week, distributed according to how the course is weighted: a theory-weighted course allocates three hours of lecture, no practical hours and six hours of independent study, while a practice-weighted course allocates three hours of lecture and six hours of practical work with no independent study. Staff workload is counted from the lecture and practical hours; independent study is not counted. Because the competence at issue is performative, comprising listening, speaking, reading and writing for professional purposes, its conventional delivery falls under the practice-weighted allocation, in which the full complement of hours is staffed.

On that basis, contact hours absorbed were calculated as the hours that classroom delivery of the same 70-hour curriculum to 128 learners would have required. The institution schedules language courses in sections of 40, so 128 learners require 4 sections, giving 4 multiplied by 70, or 280 contact hours per offering. Marking absorbed was calculated as 128 learners multiplied by 23 assessments, comprising 22 lesson tests and one final test, giving 2,944 papers. At 3 to 5 minutes per paper for a 20-item test marked by hand, a range consistent with published estimates for objective marking, this represents 147 to 245 hours.

The same framework identifies what the provision did not absorb. The online course occupies the position of a course weighted towards independent study, in which learners work at their own pace and staff involvement takes the form of coordination rather than instruction. Coordination of this kind persists: learners require enrolment, monitoring, and response to individual queries, and the course itself requires maintenance and item revision between offerings. Expressed in the terms of the framework, the provision moves the hours from the practical column to the independent-study column but does not empty the staffing requirement altogether. The hours that remain are not a residue of what automation failed to absorb; they are the facilitation the design intends. A course premised on self-directed learning requires someone to enrol learners, to notice when the record shows a learner has stalled, and to answer the questions that a system cannot anticipate, and those functions were assigned to staff by design rather than left to them by default. The estimates below therefore describe the delivery and marking functions the system took over, not the whole of what conventional provision would demand of staff.

Three assumptions underlie these figures, and each qualifies them. The section size of 40 reflects institutional practice and would not hold where sections are smaller. The marking rate assumes hand marking of objective items without item analysis; where optical scanning is already available, the marking figure would be

substantially lower. Neither figure includes the effort required to build the course, comprising the writing of 22 lessons, the construction of the lesson tests and the four achievement tests, and installation and revision on the platform across three cycles, nor the maintenance and learner-support time the provision continues to require. The study did not record either quantity in hours, and neither can therefore be entered into the estimate. The estimates are therefore of workload avoided under stated assumptions at the point of delivery, not of net institutional saving.

Instruments and analysis

Needs were assessed with a dual-response questionnaire completed by 385 pre-service teachers, in which respondents rated the desired condition and the current condition for each of 22 competency items. The modified priority needs index was computed for each item as the difference between the two ratings divided by the current condition, following Wongwanich (2019), with values from 0.20 interpreted as indicating high need. The index for writing was 0.605 against an overall value of 0.448.

Achievement was measured with four multiple-choice tests covering listening, speaking, reading and writing, each worth 20 marks and 80 in total, presenting professional scenarios and requiring selection of the response consistent with linguistic and professional principles. Tests were machine-scored. The mean item-objective congruence index was 0.805, and Cronbach's alpha ranged from 0.786 to 0.837. These instruments assess recognition of appropriate professional language and do not assess production; the outcome variable is accordingly described throughout as recognition-based knowledge of professional communication principles rather than as communication skill.

The item format was determined by the platform rather than chosen on measurement grounds. The institutional system accepts responses only as the selection of one option from a set, and provides no facility for extended text entry, audio capture or file upload against which a rubric could be applied. Assessment of speaking and writing therefore had to be constructed within that constraint, and took the form of professional situations for which the learner selects the appropriate spoken response or the appropriate written formulation. This is a defensible way of assessing whether a learner recognises what the situation calls for, and it is not an assessment of whether the learner can produce it. The constraint is worth stating plainly because it belongs to the same phenomenon the article examines: the boundary of what an automated system can absorb is set in part by what the system can accept as a response, and a platform that cannot receive production cannot assess it.

Achievement was analysed with paired-samples t-tests and Cohen's d in its average-standard-deviation form with 95% confidence intervals (Lakens, 2013). The comparison between receptive and productive lesson pass rates was tested with the Mann-Whitney U test, chosen because the lesson-level distributions are bounded at 100% and markedly skewed, with the Welch t-test reported alongside. Proportions of learners whose scores declined were compared with a chi-square test of independence. Four indicators of measurement quality were recorded in every cycle, namely the standard deviation of post-instruction scores, the correlation between pre-instruction and post-instruction scores, the proportion of learners reaching 80% of the maximum, and the number whose scores declined.

Learner reflections were analysed thematically following Braun and Clarke (2006). Two coders independently coded the full corpus of 128 responses, generating initial codes inductively and grouping them into candidate themes, which were then reviewed against the coded extracts and the full dataset. Inter-coder agreement was 0.78 as measured by Cohen's kappa. Disagreements, which concerned 14 extracts, were resolved by discussion with reference to the coding definitions, and where agreement could not be reached, the more conservative coding was adopted. Themes reported by at least a quarter of respondents were carried forward to the development panel.

RESULTS

Estimated workload absorption and the accompanying learning outcome

Classroom delivery of the same curriculum to 128 learners in 4 sections would have required 280 contact hours per offering, and hand marking of the 2,944 test papers would have required 147 to 245 hours. The provision absorbed both under the assumptions stated above, and continues to do so at each offering without further teaching hours, against a one-off development investment that the study did not quantify.

The learning outcome accompanying this provision was substantial. In cycle 2, whose measurement properties were the strongest of the three, post-instruction scores of 49.37 (SD = 8.24) exceeded pre-instruction scores of 31.66 (SD = 7.51), $t(127) = 30.03$, $p < .001$, $d = 2.25$, 95% CI [16.54, 18.87]. Against the meta-analytic benchmark of 0.71 (Morina et al., 2025), this is approximately three times the mean, a difference attributable to the one-group design, the different unit of analysis and a low pre-instruction baseline.

The educational significance of this gain should be read alongside its statistical magnitude. A post-instruction mean of 49.37 represents 61.7% of the maximum, which is a modest level of attainment in absolute terms and falls below the 65% band that the design team itself set as the minimum for competent performance. The gain is large

relative to where learners began; it does not establish that learners reached a professionally adequate standard. Reporting both figures is necessary because an effect size alone conveys the first and not the second.

Platform records show that learners spent between 429 and 1,521 minutes in the system, a mean of 1,037 minutes. The curriculum allocates 70 hours, or 4,200 minutes, which functions as the time made available rather than the time required. The learner who spent the longest used 1,521 minutes, or 36.2% of the allocation, and the cohort mean of 1,037 minutes represents 24.7% of it. No learner exceeded the allocation, and all completed all 22 lessons. The self-paced design thus produced completion well within the time budgeted, with individual variation of approximately three and a half to one in time taken, a range that no timetabled course could accommodate.

Was absorption even across the competence

Lesson-level first-attempt pass rates are recorded automatically and are independent of the achievement tests. Lessons targeting receptive skills showed a mean first-attempt pass rate of 86.8% (SD = 11.8, n = 8 lessons), with 1 lesson below 70% and the lowest at 64.1%. Lessons targeting productive skills showed 82.4% (SD = 16.1, n = 14 lessons), with 3 lessons below 70% and the lowest at 54.7%. The difference of 4.4 percentage points was not statistically significant, U = 64.0, p = .605, r = .11; the Welch t-test gave t(20) = 0.74, p = .471.

Individual change points in the same direction and is likewise not significant. Declines here are within-subject, comparing each learner's post-instruction score with their own pre-instruction score on the same skill in cycle 2. Expressing declines as a proportion of the 128 learners assessed in each skill controls for the unequal number of lessons: 5.5% declined in listening, 10.2% in reading, 6.2% in speaking and 13.3% in writing, giving 7.8% across the receptive skills and 9.8% across the productive skills, $\chi^2(1) = 0.39$, p = .532.

The direction is consistent across four indicators derived from different data sources, namely mean pass rate, lowest lesson, lessons below threshold and proportion declining. None reaches significance. These indicators are the closest approach to the modality question that the available data permit, since no measure of production was administered; they concern the difficulty learners encountered in lessons addressing productive skills, not their capacity to produce. The distinction matters for how much weight the pattern can carry, and we take it to warrant a hypothesis and no more. With 8 and 14 lessons, the comparison had power of approximately .13 to detect a difference of the observed magnitude, so a real difference of this size would not be expected to reach significance in this design. The pattern is consistent with the theoretical expectation set out above but does not establish it, and we report it accordingly.

Three considerations further qualify the lesson-level comparison. The lesson groups differ in number, 8 against 14. They also differ in position within the course, since the sequence proceeds from receptive to productive, so productive lessons were encountered later and by learners who had already completed the receptive material. And they differ in the content and cognitive demand of the material itself, which the pass rate cannot separate from modality. The comparison is therefore not a controlled contrast between modes but an observation across lessons that vary in several respects at once.

One further caution concerns the description of the lesson-level indicator. Six of the 22 lessons recorded first-attempt pass rates of 100%, so the indicator is not free of ceiling effects, as an earlier formulation of this work stated. It varies more widely than the achievement tests, ranging from 54.7% to 100%, and that wider variation is what makes it informative; it is not unbounded.

Convergence with the needs analysis

The most demanding lesson in the course was official correspondence writing, passed on the first attempt by 54.7% of learners. The competency item carrying the highest needs index of the 22 assessed more than a year earlier, at 0.605, concerned official correspondence. The two datasets differ in kind, the first being behaviour recorded automatically during delivery and the second stakeholder perception gathered before development began, and their agreement on the same point is a form of triangulation. It is a convergence between two indicators and not a test of the modality hypothesis, since both concern writing specifically rather than productive skills generally.

What learners reported

Reflection data identified the operation of the system as the most frequently reported obstacle in the second cycle, mentioned by 41% of respondents. One learner wrote that the number of steps needed to reach a test was discouraging: "I had to go through several screens before the test appeared, and once the connection dropped I had to begin again." Another described the loss of submitted work: "When I left the system and came back, my answers were gone and I had to do them again in the time that was left." Training in operating the system was provided before cycle 3 in response.

A second theme, reported by 34% of respondents, concerned the assessment of productive skills. One learner wrote: "In the speaking lessons I choose the right answer, but I do not actually speak." Another observed of the writing lessons: "I know which sentence is correct, but writing the whole letter myself is different." These accounts are consistent with the construct under-representation

described in the introduction, though they are learner perceptions of the assessment rather than evidence about their competence.

DISCUSSION

Workload substitution as the useful question

The results support a claim about what automated provision did in this setting. A competence mandated by regulation, for which the curriculum had no space and the faculty no staff time, was delivered to 128 learners with substantial measured improvement in recognition-based knowledge, against an estimated 280 contact hours and 147 to 245 marking hours that conventional delivery would have required. For institutions facing mandates without resources, this is a consequential quantity, and expressing it in hours rather than in general terms allows it to enter institutional planning.

A further element of conventional provision, which the estimates do not attempt to quantify, is the time absorbed by assembling a class at all. In a small rural institution learners travel independently by motorcycle or bicycle, and a timetabled session begins when enough of them have arrived. The self-paced design removes this constraint entirely, as the platform records show: learners spent between 429 and 1,521 minutes on the course, a range of approximately three and a half to one, and completed at times of their own choosing. No timetabled arrangement could accommodate that variation, and the coordination it would demand of staff is real but was not measured here.

The claim requires one clarification before the qualifications, because the setting makes the counterfactual unusual. The faculty had never delivered this competence as a taught course, having neither the credit space nor the staffing to do so, and the 280 hours therefore do not represent staff time that was previously committed and has now been freed. They represent the price that conventional delivery would exact under the institution's own workload framework, and it is precisely because that price could not be met that the competence went undeveloped despite being mandated. What the provision did was not to release a budget that existed but to bring within reach a requirement that had stood beyond it. Readers in better-resourced settings should read the figure accordingly: where the course would otherwise have been taught, the same estimate describes hours that could be redeployed; where it would not, as here, it describes the cost of an obligation that would otherwise have remained unmet.

Three qualifications attach to the claim. The figures are estimates against a counterfactual that was not observed. They exclude the development investment, which the study did not record in hours, and the continuing maintenance and support the provision requires. And workload avoided is not capacity released unless an institution acts to redeploy it; where staff hours are simply not scheduled, the avoided workload registers as an absence rather than as a resource. We therefore describe the quantity as estimated workload substitution and reserve the language of released capacity for cases where redeployment is documented.

The recognition-production boundary as a hypothesis

The study set out to test whether absorption was even across the competence and found a consistent direction without statistical support. Productive lessons were passed at first attempt less often than receptive lessons, contained more lessons below threshold and were associated with more within-subject decline, but no comparison reached significance, and the lesson-level test was underpowered. We therefore advance the recognition-production boundary as a hypothesis for direct test rather than as a finding of this study.

Testing it properly requires what this study lacked, namely a measure of production, and the reason it was lacking is itself part of the argument: the platform accepts only selected responses and provides no facility for extended text, audio or file submission, so a production measure could not have been administered through the same system that delivered the course. An institution wishing to assess production alongside automated delivery must therefore either adopt a platform that accepts it or retain the assessment with staff, which returns the question to where workload should sit. A design capable of the test would also benefit from a comparison group, and the setting affords one that raises no ethical difficulty. Learners who enrol voluntarily but await a later offering constitute a waitlist against which a cohort receiving the provision can be compared, without withholding from anyone a course they have asked for; a stepped introduction across subject majors would serve the same purpose while ensuring that every major receives the provision within the academic year. A design capable of the test would assess the same learners on recognition items and on a production task judged by trained raters, and compare gains across the two. Such a design would also permit the alternative explanations that the present data cannot exclude. The first is instructional: productive lessons may have been taught less effectively rather than assessed less adequately, since the course afforded fewer opportunities for extended production than for guided recognition. The second is sequential: productive lessons came later in the course, when learner engagement may have declined, and the design confounds modality with position. The third is a measurement artefact: if productive items were harder in ways unrelated to modality, the same pattern would follow. We regard the assessment explanation as the most consistent with the learner reflections and with the cognitive literature (Anderson, 1982; DeKeyser, 2015), but the study does not adjudicate between them.

What the study does establish is narrower and worth stating precisely. Learners improved substantially in recognising appropriate professional language. Whether they improved in producing it is not addressed by these data, and the effect size reported here should not be read as an estimate of gains in communication skill.

Redirecting absorbed workload

If absorption is uneven in the direction observed, the practical consequence concerns where staff attention should go. The question is sharper in a course designed around self-direction, because the facilitation role it establishes is already the place where staff time is concentrated, and adding to it is therefore a matter of extending an existing function rather than inventing a new one. Four uses of avoided workload follow from the results and from the learner reflections. The first is assessment of production: staff time released from marking objective items could be applied to reading and responding to a small number of authentic written products, for instance one official letter and one post-teaching reflection per learner per term. The second is targeted intervention in the lessons the platform identifies as most demanding, which in this course were official correspondence writing and constructive speaking. The third is the provision of oral feedback, which no component of the automated course supplied. The fourth is course maintenance and item revision, which the estimates above exclude and which the ceiling observed in later cycles shows to be necessary.

These are proposals rather than findings. Whether redirecting workload in these ways improves outcomes is a question the present design cannot answer and one we recommend for study.

Conditions of transfer

Three features of the setting condition the transfer of these findings, and one observation bears on relevance. The 117 pre-service teachers who enrolled voluntarily outside the trial, under no obligation and with no measurement attached, indicate that the provision answered a need the population recognised for itself, although nothing is known about what they did with it. Relevance of this kind is a precondition of the workload argument, since provision that learners do not take up absorbs no workload, however efficiently it is designed. The estimates assume section sizes of 40 and hand marking of objective items; institutions with smaller sections or existing optical marking would obtain smaller figures, and the method of estimation rather than the figures is what transfers. The diglossic context means learners began with extensive receptive familiarity and limited productive occasion, so the receptive-productive asymmetry may be smaller in settings without this feature. And the competence in question is one for which authentic professional documents are readily specified; competences with less determinate output would be less amenable to the lesson structure used here. What we expect to transfer is the distinction between estimated substitution and realised saving, and the practice of examining absorption separately for recognition-based and production-based components rather than for a course as a whole. Within the institution, the intended route is scale rather than replication: the design principles derived here are to be applied to a full twelve-credit programme, for which the estimates reported would rise in proportion to the hours involved while the boundary between what a platform can absorb and what facilitation requires would remain where this study locates it.

Limitations

The one-group design precludes causal attribution, and gains cannot be separated from maturation, history or repeated testing. The same cohort completed three cycles, so cycle 3 results reflect cumulative exposure, and its pre-instruction scores are not a baseline. The field trial sample is a single cohort of pre-service teachers at one institution, and the formative tryout that preceded it drew on a different population, so the two phases are not directly comparable. All assessments used multiple-choice items, so the study measures recognition and is silent on production; this is the most consequential limitation and the reason the modality comparison is offered as a hypothesis. The workload figures are estimates against an unobserved counterfactual and exclude development, maintenance and support. The lesson-level comparison confounds modality with sequence position and content, and was underpowered. Reflection data were gathered in writing and anonymously, which may favour concrete concerns over diffuse ones. The study was conducted at a single institution with one platform. Finally, the first author designed, taught and evaluated the course, although machine scoring removed rater bias from the achievement measures.

Conclusion

An online course delivered without contact hours met a regulatory requirement that the institution had no capacity to meet conventionally, producing substantial improvement in recognition-based knowledge among 128 pre-service teachers, all of whom completed the course within the time allocated, and absorbing an estimated 280 contact hours and 147 to 245 marking hours per offering. Lessons targeting productive skills were passed less often on the first attempt than those targeting receptive skills, but the difference was not significant and is offered as a hypothesis for direct testing rather than as a finding. Three

contributions are offered, and it is worth stating them separately because the claims differ in strength. The first is a method: workload substitution can be estimated against an institution's own framework rather than asserted, and the estimate can be disaggregated by function so that what a platform absorbs is distinguished from what facilitation retains. The second is an outcome: learners improved substantially in recognising appropriate professional language under conditions where no taught provision had previously existed, with the limits of that measure stated. The third is a hypothesis: that automated provision serves recognition better than production, for which this study offers converging but non-significant indications and a design capable of testing them. Two distinctions deserve to survive this study: between workload estimated to be avoided and capacity actually released, and between recognition of appropriate professional language and the production of it. A third is implicit in the design: the hours that remain with staff in a self-directed course are the facilitation the design intends, not the remainder of a substitution that fell short. Institutions adopting automated provision under resource constraint would do well to hold both.

DECLARATIONS

Ethics approval and consent to participate

The study was approved by the Human Research Ethics Committee of Kalasin University, certificate of approval number HS-KSU 051/2567, issued on 19 November 2024. Written informed consent to participate was obtained from all participants. Participation was voluntary, and no participant received academic credit or penalty in connection with the research.

Consent for publication

No individual participant is identifiable in this article. All learner data are reported in aggregate or under anonymised identifiers, and the quotations reproduced above are presented without any information that would permit identification.

Availability of data

The datasets analysed in this study, including lesson-level platform records, the anonymised achievement data and the coding frame used in the thematic analysis, are available from the corresponding author on reasonable request, subject to the conditions of the ethics approval under which they were collected.

Competing interests

The authors declare that they have no competing interests.

The online course reported here was developed within the authors' institution and is not offered on a commercial basis.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

ACKNOWLEDGEMENTS

The authors thank the administrators, lecturers and pre-service teachers of the participating faculty for their contribution to the needs analysis, the design workshops and the field trial, and the experts who reviewed the instruments and the innovation. We are grateful to the two anonymous reviewers, whose comments materially improved the reporting and the claims of this article.

REFERENCES

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369-406.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115.

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Campbell, D. T., and Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

DeKeyser, R. (2015). *Skill acquisition theory.* In B. VanPatten and J. Williams (Eds.), Theories in Second Language Acquisition: An Introduction (2nd Ed., pp. 94-112). New York: Routledge.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning and Assessment, 5*(1), 1-35.

Ferguson, C. A. (1959). Diglossia. *Word, 15*(2), 325-340.

Fleckenstein, J., Liebenow, L. W., and Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6,* 1162454.

Knowles, M. S. (1975). *Self-Directed Learning: A Guide for Learners and Teachers.* New York: Association Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4,* 863.

McKenney, S., and Reeves, T. C. (2019). *Conducting Educational Design Research.* 2nd Ed. London: Routledge.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Morina, F., Fütterer, T., Hübner, N., Zitzmann, S., and Fischer, C. (2025). Effects of online teacher professional development on teacher, classroom and student level outcomes: A meta-analysis. *Computers and Education, 228,* 105247.

UNESCO (2023). *Global Report on Teachers: Addressing Teacher Shortages.* Paris: United Nations Educational, Scientific and Cultural Organization.

Warschauer, M., and Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*(1), 22-36.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wongwanich, S. (2019). *Needs Assessment Research.* 4th Ed. Bangkok: Chulalongkorn University Press.