International Journal of Chinese Language Teaching (2026)
Vol.7 (3) 133-152 https://doi.org/10.46451/ijclt.260411

Article

Constructing a Data-driven Language Proficiency Pathway: An Intervention with an HSK-4 Intelligent System

Nanxi Bian
University of Macau, China

Qingyu Gao*
Shenzhen MSU-BIT University, China

Received: 15 December 2025/Accepted: 8 March 2026/Published: 10 April 2026

Abstract
Traditional "one-size-fits-all" HSK preparation models may constrain proficiency development by insufficiently addressing diverse learner needs. This study investigates the efficacy of a data-driven intervention for HSK Level 4 implemented through an intelligent practice system. Grounded in SLA theories, the system uses big data analytics to analyse learner errors, weaknesses, and behaviours. It then dynamically generates personalised "i+1" exercises for HSK Level 4's core item types, enabling precise learning interventions. An 8-week quasi-experiment with 49 Russian-speaking students employed a natural usage-based grouping to examine the association between system engagement and learning outcomes. Quantitative analysis revealed a strong positive correlation between system usage intensity and score gains (r = .811, p < .001). This study provides empirical evidence suggesting that big data intervention may help address the limits of homogenised instruction. By enabling real-time tracking, diagnosis, and personalised feedback, it outlines a data-informed pathway that may inform HSK preparation practices and ongoing digital transformation efforts in language education.

Keywords
Data-driven language learning, HSK preparation, intelligent tutoring system, second language acquisition, personalised learning

1 Introduction

In an era of deep integration between globalisation and educational digitalisation, International Chinese Language Education is accelerating its transformation from being experience-driven to data-driven. Data is increasingly regarded as a key element in addressing challenges of pedagogical homogenisation and achieving the precise development of language proficiency. The Action Plan for Online International Chinese Language Education (2021-2025) identifies "data-driven teaching reform" as a strategic priority, proposing the comprehensive construction of a digitalised and intelligent Chinese language education

*Corresponding author. Email: gaoqy219@gmail.com

ecosystem by 2025, thereby highlighting the strategic value of data in enhancing pedagogical efficacy. The Chinese Proficiency Test (HSK) Level 4, as a key benchmark for certifying the Chinese proficiency of international students, serves as both an entry requirement for advanced academic pursuits and an important credential for professional development. HSK Level 4 was selected as the intervention context for both theoretical and pedagogical considerations. First, it represents an important transition phase from intermediate to upper-intermediate proficiency, where learners face a substantial increase in lexical (2,500 words) and grammatical (120 points) complexity. Second, the diverse item types (e.g., sentence sequencing, picture-cued writing) demand differentiated skills, creating a heterogeneous landscape of learner weaknesses that is both challenging for traditional one-size-fits-all instruction and potentially well-suited for data-driven diagnosis and personalised remediation. Moreover, pragmatic considerations related to the university curriculum offerings further informed this choice.

The traditional HSK Level 4 preparation model, constrained by uniform progress, content, and assessment, faces limitations in achieving instructional precision. This flaw originates in the absence of a data-driven mechanism throughout the instructional process. In this study, "data-driven" refers to a pedagogical approach that grounds instructional decisions in the real-time, individual-specific data rather than instructor intuition or one-size-fits-all curricular planning. This approach relies on the systematic collection, analysis, and utilisation of behavioural data (e.g., response time, practice frequency), error data (e.g., error types, mistake recurrence), and proficiency data (e.g., mastery status) to inform personalised learning interventions.

The consequences of limited data integration are observable in practical teaching scenarios. Consider, as an illustrative case, teaching practice data from an intensive HSK Level 4 course at a Chinese university. Within a limited 8 teaching-hours per week, instructors were required to address the collective needs of 49 learners without the means to systematically collect or analyse individual learning behaviours, error patterns, and competency gaps. Consequently, they could not obtain data-informed insights into learner heterogeneity regarding vocabulary mastery depth, grammatical application ability, or sensitivity to different item types. From the perspective of core data thresholds in language acquisition, prior vocabulary research suggests that lexical consolidation typically requires multiple repeated encounters, commonly estimated at six or more meaningful exposures (Saragi et al., 1978). In the traditional teaching model, however, the lack of a data-driven dynamic adjustment mechanism means the average weekly recurrence rate of HSK Level 4 core vocabulary is merely 2.7 times, considerably below commonly suggested exposure thresholds, which directly affects the consolidation of vocabulary memory. In terms of exercise allocation, the absence of data-supported learner diagnosis forces instructors to adopt a "one-size-fits-all" approach, failing to accurately allocate practice volume based on learners' weaker item types (e.g., sentence sequencing, picture-cued writing). This results in excessive practice on areas of strength and insufficient reinforcement of weaknesses. The lack of data-driven feedback is even more pronounced: the average feedback interval for a single paper-based exercise is 4 minutes, and the feedback cycle for homework can exceed one day. Furthermore, feedback is limited to superficial judgments of "right or wrong", failing to use data mining to label error types (e.g., grammatical errors, lexical collocation errors, logical errors), pinpoint proficiency weaknesses, or provide personalised remedial suggestions. This limits learners' opportunity for precise error correction based on immediate data feedback. This "no data collection, no data diagnosis, no data intervention" teaching model stands in contrast to the language education philosophy of the data-driven era and struggles to achieve precision and efficiency in language proficiency development.

With the increasing integration of artificial intelligence and educational technology, various digital tools for HSK preparation, such as "Chinese Mock Test Platforms", "HSK Mock", and "SuperTest", have emerged. However, these products have yet to overcome the core limitation that digitalisation is not equivalent to datafication, and they have not truly constructed a data-driven pathway for language proficiency development. These shortcomings manifest in three primary areas of data deficiency. Firstly, exercise allocation lacks data support. Most tools do not conduct proficiency diagnosis by mining

learners' response data, instead randomly delivering items from a fixed item bank. They are unable to accurately adjust the proportion of practice based on a learner's weak item types (e.g., grammar-based, expression-based), leading to a low correspondence between practice content and individual needs. Secondly, difficulty adaptation is not informed by data-driven models. The "*i*+1" principle from Krashen's (1982, 1985) Input Hypothesis is not deeply integrated with learner proficiency data. There is a lack of dynamic tracking and analysis of multi-dimensional data such as learners' accuracy rates, error frequencies, and practice progress. This often leads to problems of frustration from excessive difficulty or ineffective practice due to insufficient challenge, which contravenes the principle of data-driven precision adaptation. Thirdly, remedial guidance lacks data support. Feedback is limited to providing the correct answer, without employing data-driven methods like error data classification and test point correlation analysis to clarify for learners their weaknesses, the required practice volume, and effective practice strategies. This fails to meet the core need for precise remediation. These limitations suggest that many existing tools remain primarily at the stage of digital delivery, without fully leveraging the diagnostic and adaptive potential of learner-generated data, thus being unable to fundamentally resolve the inefficiency of traditional teaching.

Against this backdrop, this study integrates effective big data intervention throughout the HSK Level 4 preparation process from the core perspective of a "data-driven language proficiency development pathway". Through a closed-loop mechanism of data collection, diagnosis, and intervention, it aims to achieve precision and efficiency in language learning, with the associated intelligent practice system serving as a practical vehicle for this data-driven approach. The theoretical contributions of this study focus on the deep coupling of data, theory, and practice. Firstly, it refines a data-driven theoretical fusion framework for Second Language Acquisition (SLA). While existing research often discusses SLA theories and technological tools separately, this study uses data-driven principles as a central hub to deeply integrate three core mechanisms—the "*i*+1" difficulty adaptation from Krashen's (1982, 1985) Input Hypothesis, the "output-correction" closed loop from Long's (1996) Interaction Hypothesis, and the "focus on linguistic form" from Schmidt's (1990) Noticing Hypothesis—with the collection and analysis of learning behaviour data, error data, and proficiency data. This constructs a complete theoretical model of "learner data collection—proficiency data diagnosis—personalised intervention implementation—effect data validation", aiming to fill a gap in the field of International Chinese Language Education where theory, data, and practice are often disconnected. While preliminary efforts have explored individual components of this integration, a systematic operationalisation of SLA mechanisms within a unified data-processing architecture remains underdeveloped. Secondly, it seeks to construct a data-based theory for precise intervention in HSK Level 4. By mining learners' response data, it identifies the characteristics of weaknesses across different proficiency dimensions (vocabulary, grammar, expression) and different item types (closed-ended, semi-open, open-ended), forming an intervention logic of "data diagnosis—exercise allocation—remedial reinforcement". That is, when data identifies a learner's error rate on a certain item type exceeds 60%, the system automatically increases the practice volume for that type and matches it with targeted test point training, providing a data-driven theoretical reference for precise intervention in language teaching.

From a pedagogical practice perspective, the applied value of this study is concentrated in data-driven empowerment across all scenarios. Firstly, it provides data support for teachers' precision teaching. Through data visualisation reports, teachers can intuitively grasp the overall class's weak item types and high-frequency error points, as well as individual learners' proficiency improvement curves. Without manual statistical analysis, they can accurately identify teaching priorities, achieving data-driven precision instruction and personalised tutoring. At the same time, data can replace manual labour for repetitive tasks such as learner diagnosis and exercise selection, significantly reducing the teacher's workload. Secondly, it constructs a data-driven pathway for learners' efficient preparation. Through real-time collection and analysis of multi-dimensional learning data, the system can accurately identify a learner's weak areas (e.g., grammatical logic errors in "sentence sequencing" item, or insufficient

ability to locate details in reading comprehension). Based on data models, it dynamically adjusts exercise allocation—for instance, increasing practice volume by 30% for item types with an error rate over 50% and reducing repetitive practice for types with a mastery rate above 80%—ensuring that effort is optimally directed. Concurrently, the recurrence frequency of core vocabulary is optimised through a data model, increasing from 2.7 times in the traditional model to 7.3 times, reaching the effective threshold for language acquisition. The re-practice rate for incorrect closed-ended items reaches 68%, resulting in a higher rate of targeted reinforcement for incorrectly answered items, which may contribute to more focused remediation. Thirdly, it provides a data-driven paradigm for the digital transformation of International Chinese Language Education. The closed-loop mechanism of "data collection—diagnosis—intervention—feedback" constructed in this study can be transferred to various language test preparation and daily teaching contexts, promoting the transformation of language education from being experience-driven to data-driven and providing a replicable and scalable practical pathway for personalised language proficiency development.

In light of the above rationale, this study addresses the following research question:

To what extent is the intensity of engagement with a data-driven intelligent practice system associated with improvements in HSK Level 4 proficiency?

Additionally, does a graded, dose-response pattern emerge across different levels of system engagement?

2 Literature Review

2.1 The lag between digitalisation and datafication in international Chinese language education

Driven by the global wave of educational digitalisation, International Chinese Language Education is undergoing a paradigm shift from being primarily offline to being digitally empowered. However, the field is often characterised by an emphasis on digitalisation without a corresponding deep integration of data-driven mechanisms. Research both domestically and internationally has explored the integration of intelligent technology and language teaching, but relatively few studies have systematically addressed the limitation of emphasising tool digitalisation while underutilising learner-generated data, thereby limiting the construction of a fully data-driven teaching and learning ecosystem.

The digital transformation of International Chinese Language Education in China is primarily policy-driven. The Action Plan for Online International Chinese Language Education (2021-2025), issued in 2021, clearly sets the development goal of "basically achieving digitalisation, intelligence, and ubiquity by 2025", listing the development of intelligent tools as a core task. The 2023 statement that "educational digitalisation is a crucial breakthrough for opening up new tracks in educational development" has further accelerated the application of digital technology in this field. At the level of research and practice, scholars have focused on the supplementary value of digital tools to traditional teaching. Li (2020) pointed out that after 2020, digital and intelligent resources such as MOOCs and adaptive learning systems have become important vehicles for breaking through the time and space constraints of the classroom, achieving the digital presentation and dissemination of teaching content. Yuan and Wu (2023) confirmed through empirical research that generative AI technology can optimise learners' autonomous learning models and provide digital solutions for personalised practice. Addressing the challenges of traditional classrooms, such as the difficulty of accommodating multiple native language backgrounds in mixed-nationality classes and delayed feedback, Wang and Zhang (2024) proposed the need to develop digital tools tailored to specific population characteristics. Hu and Zhang (2023) also emphasised that teachers urgently need digital and intelligent tools for precise and personalised teaching to compensate for the shortcomings of limited class hours. However, existing explorations in China largely focus on the functional implementation of digital tools. Many existing studies primarily focus on the effects of technology application, with comparatively less attention to the systematic design of data

collection, analysis, and intervention processes. The integration between SLA theory and data-informed instructional design remains relatively underdeveloped. Although some pilot projects have basic digital functions, they have not established a correlation model among learner behaviour data, error data, and proficiency development. As a result, in some cases, "personalisation" remains limited to surface-level differentiation rather than dynamic, data-driven adaptation, thus failing to resolve the core contradictions of traditional teaching. Xinhua News Agency (2024) reported that on the digital transformation of university English education, a common problem in current educational digitalisation is the emphasis on technology application over data-driven approaches, and the field of International Chinese Language Education similarly lacks a pedagogical reconstruction centred on data as a core element.

Internationally, research on intelligent second language teaching has a longer history, forming a research framework that combines technology with SLA theory. However, research specifically on Chinese as a second language has yet to fully explore the potential of datafication-oriented approaches. At the technological level, Natural Language Processing and adaptive learning systems have been widely integrated into language education. The concept of "digital assessment literacy" proposed by Eyal (2012) provides a theoretical basis for using intelligent tools to assist in learner assessment. A bibliometric analysis by Liang et al. (2021) also shows that personalised learning path generation and learning behaviour data analysis have become research hotspots, and the effectiveness of adaptive systems in vocabulary and grammar teaching has been empirically supported. In terms of theoretical integration, Krashen's (1982, 1985) Input Hypothesis has become a core design principle. Some mainstream intelligent language systems have attempted to achieve "i+1" difficulty adaptation through technology. For example, the Deep English platform evaluated by Meng and Zhang (2024) can adapt to learners' levels through features like speed adjustment and material grading. Zhang et al. (2024) also confirmed that the personalised progress control of intelligent systems can reduce classroom anxiety and compensate for the inability of traditional teaching to accommodate individual differences. However, international research has two main limitations. Firstly, most findings focus on major languages like English and Spanish, with a lack of targeted research on Chinese. The unique item structures and vocabulary-grammar systems of HSK Level 4 may not be fully accommodated by data models originally developed for alphabetic languages. Secondly, the depth of data application remains uneven across contexts. Although most systems can collect basic learning data, they have not achieved a closed loop of data diagnosis, practice optimisation, and effect validation. For instance, the classification of error data often remains relatively surface-level and insufficiently connected to test-point logic or acquisition patterns. This may limit the potential of data to address core needs such as the precise reinforcement of weak item types and informed allocation of practice volume, failing to realise the concept of "data-driven fine-grained decomposition of learning units" advocated by the Global Talk International Chinese Language Education Intelligent Body.

2.2 The core problem in current international Chinese language education: The dilemma of datafication deficiency masked by digitalisation

A synthesis of existing research and pedagogical practice reveals that a central challenge facing International Chinese Language Education in its digital transformation is the lack of precision in intervention, caused by insufficient datafication capability. This manifests in three major deficiencies, the root of which is the failure to transform digital tools into data-driven teaching capabilities.

First, the difficulty in meeting personalised learning needs stems from the absence of learner data collection and analysis. Traditional International Chinese Language classrooms are often mixed-nationality and mixed-language-family settings. The 49 learners in this study's context have significant differences in their native language backgrounds and Chinese proficiency levels. However, digital tools have often only achieved the online presentation of teaching content, without establishing a multi-dimensional learner data collection mechanism. There is a lack of precise capture of both proficiency

data, such as vocabulary mastery rates and grammar application error rates, and behavioural data, such as response times and practice frequencies. This makes it difficult for teachers to gain data-informed insights into individual differences. Teachers can only conduct centralised teaching based on the "average level", leading to a dilemma where high-proficiency learners engage in ineffective repetitive practice and low-proficiency learners face input beyond their cognitive scope. This situation appears to diverge from the principle of providing appropriately calibrated input as suggested by Krashen's (1982, 1985) Input Hypothesis and impedes the implementation of precise feedback as advocated by Hu and Zhang (2023). This "digitalisation without data support", even when relying on digital platforms, has not escaped the shackles of the traditional "one-size-fits-all" model.

Second, the difficulty in accommodating fragmented learning needs stems from the lack of data-driven exercise module design. The after-class study time of non-Chinese major international students is often fragmented. However, existing digital tools simply replicate traditional practice models without optimising the practice structure based on learning behaviour data. From a data threshold perspective, HSK Level 4 core vocabulary requires an average effective recurrence of over six times per week to be firmly mastered (Saragi et al., 1978). However, traditional digital exercises lack a data analysis and dynamic adjustment mechanism for recurrence frequency, resulting in an average weekly recurrence rate of core vocabulary of only 2.7 times, considerably below commonly suggested exposure thresholds. At the same time, the duration of exercises is not adapted to fragmented scenarios. The inability to break down learning units through data models makes it difficult for learners to complete effective practice in their spare moments. The ubiquity advantage of digital tools has not been translated into learning efficacy through data-driven design.

Third, the difficulty in implementing immediate feedback stems from the insufficient in-depth mining of error data. Timely feedback is a critical component of second language acquisition. However, the feedback mechanism of existing digital tools remains at the level of "digital marking", without achieving "data-driven analysis". The feedback interval for traditional paper-based exercises ranges from 4 minutes to 1 day. Although digital tools have shortened the feedback time, the content is still dominated by "right or wrong" judgments, lacking in-depth mining of error data. It does not label error types (e.g., lexical collocation errors, grammatical logic errors) through data, trace them back to associated test points, or provide targeted analysis based on the learner's native language background. Such feedback, when not supported by structured data analysis, may be insufficient to help learners fully understand the nature of their errors, how to correct the error, and how much to practise. Even with immediate feedback, the core goal of error correction and reinforcement is difficult to achieve, falling far short of the concept of a "data-driven diversified evaluation system" proposed by Xinhua News Agency (2024).

2.3 Research gap and study rationale

In summary, while existing research has laid a foundation for the digital transformation of International Chinese Language Education, the transition from mere digitalisation to meaningful datafication remains a continuing and not yet fully resolved area of development and scholarly debate. Existing research focuses on the functional implementation of intelligent tools, lacking a systematic design of data-driven mechanisms. It pays attention to the effects of technology application but has not constructed a complete closed loop of data collection, diagnosis, intervention, and validation. Although personalised teaching is frequently discussed, systematic data mining approaches for determining practice allocation and reinforcement pathways remain relatively limited. This tendency to emphasise digitalisation while underemphasising datafication may contribute to the persistence of problems in HSK Level 4 preparation, such as insufficient vocabulary recurrence, imbalanced reinforcement of item types, and low precision of feedback.

Therefore, this study aims to address this gap by focusing on datafication empowerment. Taking HSK Level 4 as a specific scenario, it constructs a data-driven language proficiency development

pathway. By collecting multi-dimensional information such as vocabulary mastery data, error type data, and learning behaviour data, it establishes a closed-loop mechanism of learner data diagnosis, precise exercise allocation, and effect data validation. This seeks to transform digital tools into vehicles for data-driven precision intervention, with the aim of addressing the three major challenges of personalisation, fragmentation, and immediate feedback, and providing a datafication-focused upgrade solution for the digital transformation of International Chinese Language Education.

3 Construction of a Data-Driven Dynamic Learning System

To investigate the feasibility and impact of a data-driven proficiency pathway, we developed an intelligent practice system as the research vehicle. Its architecture was designed with data flow as the central principle, comprising three layers of front-end data collection, middle-end data processing, and back-end data storage and iteration. Through full-process data capture, real-time analysis, and dynamic application, it is designed to support automated item generation, personalised practice guidance, and dynamic optimisation of the learning pathway, forming a closed-loop mechanism of data collection, analysis, intervention, and update. The primary design goal was to ensure that every core function—item generation, feedback, path adjustment—was explicitly governed by the analysis of learner data, thereby operationalising the hypothesised data-driven intervention model for empirical examination.

3.1 Operationalising SLA theories in the data-driven system

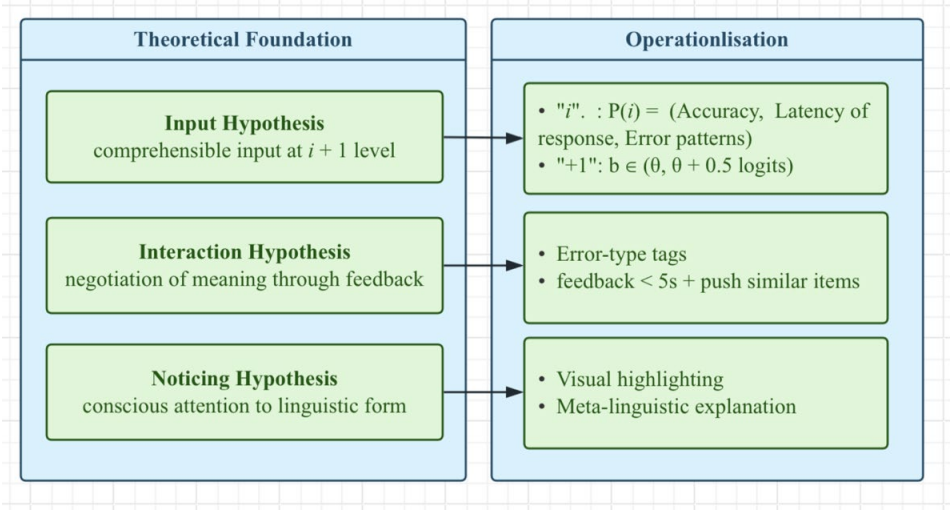
To ground key SLA theories in data-driven interventions, this study integrates three core theoretical mechanisms, i.e., Krashen's (1982, 1985) Input Hypothesis, Long's (1996) Interaction Hypothesis, and Schmidt's (1990) Noticing Hypothesis, into a coherent data-processing pipeline (see Figure 1). This integration aims to ensure that theoretical principles are deeply embedded within the data logic and operational architecture of the system.

Specifically, the learner's current proficiency (i) is operationalised as a composite profile, derived from a dynamic vector $P(i) = (A, L, E)$. A represents the weighted average accuracy rate. L denotes the normalised average response latency (time per item), inversely correlated with automaticity. E captures the distribution of error patterns across major categories (e.g., grammatical vs. lexical). This vector is continuously updated in the back-end's learner-knowledge point matrix. The optimal challenge ("+1") is designed as an algorithmic difficulty buffer. The system maintains an estimate of the learner's ability (θ) and, when generating exercises, targets items with a calibrated difficulty parameter (b) satisfying: $\theta < b < \theta + 0.5$. The lower bound ($\theta$) ensures the task is not too easy, while the upper bound ($\theta + 0.5$ logits), a buffer zone calibrated during our pilot study to maintain engagement, prevents excessive frustration. This bounded difficulty range operationalises the Input Hypothesis by translating the abstract construct of "i+1" into a measurable selection rule within the item generation algorithm.

The system algorithmically approximates certain elements of the negotiation and repair sequences described in the Interaction Hypothesis. Upon answer submission, it provides immediate feedback (< 5 seconds) tagged with specific error types generated by the knowledge point recognition model. Subsequently, the system automatically pushes reinforcement exercises containing the same or similar linguistic forms, thereby approximating the interactional modification cycle described in the Interaction Hypothesis, in which feedback and reformulation facilitate comprehensible input and pushed output. The Noticing Hypothesis is realised through in-depth data parsing and targeted presentation. The system categorises error data and links it precisely to specific linguistic forms. In the feedback interface, learner attention is consciously guided towards the gap between their output and the target form through visual cues, meta-linguistic explanations, and linked knowledge point recommendations. In this way, attentional orientation is systematically regulated by algorithmic error tagging (e.g., conjunction word order error) rather than incidental exposure, thereby operationalising noticing as a data-triggered cognitive event.

Figure 1

Framework Driven by Theory-Data Integration

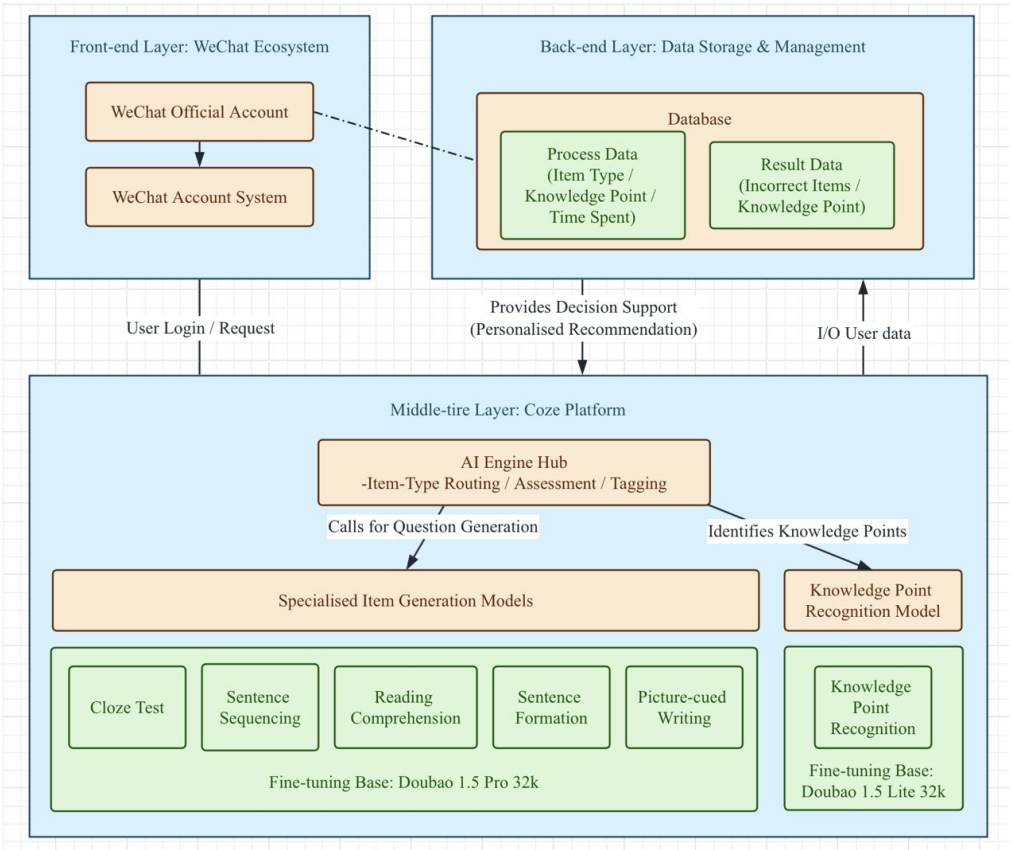


3.2 Overall system architecture: A layered design for data integration

The system adopts a layered architecture with data as its core operational element, comprising a front-end data collection layer, a middle-tier intelligent processing layer, and a back-end data modelling layer (see Figure 2). This design ensures the continuous collection, real-time processing, and dynamic application of data throughout the entire workflow from learner interaction to personalised intervention, providing the foundational support for a theoretically informed, data-driven closed loop.

Figure 2

A Layered Architecture with Data



3.2.1 Front-end layer: Unified data acquisition portal

The front-end employs a WeChat Official Account as a unified portal for interaction and data acquisition. Its design adheres to the principle of "low barrier to entry, high fidelity of capture," aiming to minimise user friction while improving the continuity and granularity of learning process data collection. Given the widespread use of WeChat among international students in China (Ministry of Education, 2022), the platform was selected as the unified access portal. Its familiarity reduces the barrier to entry and supports the continuity of long-term behavioural data collection.

The core function of the front-end transcends the conventional presentation of exercises; it is essentially a multi-dimensional learning behaviour recorder. Through the interactive interface, the system captures three primary streams of data in real time. Firstly, it logs basic operational data, including login frequency and timing, the duration of individual practice sessions, and the combination of item types selected for each session. Secondly, it captures response process data, encompassing the time spent on each item, the number of answer modifications during the attempt, the final submission timestamp, and hover times over different options. Thirdly, it records feedback interaction data, such as answer correctness, the duration spent viewing provided explanations for incorrect answers, clicks on example sentences, and whether a repeat practice was actively initiated. All raw data, including unstructured logs and structured response records, are synchronised in real time to the back-end database. Furthermore, the front-end's interactive UI design strictly adheres to the specifications of the five core item types in the HSK Level 4 exam (e.g., cloze, sentence ordering, picture-cued writing), ensuring that the specific competency dimension data corresponding to each item type is captured accurately, thereby preventing any data omission due to interface incompatibility.

3.2.2 Middle-tier layer: The data-driven intelligent processing core

The middle-tier is constructed on an AI agent orchestration platform (Coze), chosen for its compatibility with collaborative AI agent workflow design. It consists of a cluster of intelligent agents that function as the system's "data analysis and decision engine." This layer comprises six specialised models, whose core design philosophy is to enable adaptive transformation from data to instructional intervention directives, rather than executing static, pre-defined functions.

The Automated Item Generation Model Set consists of five specific generation modules, each configured to produce one of the five core HSK Level 4 item types. These modules are built upon a large-scale transformer-based language model (Doubao 1.5 Pro 32k) and guided by structured prompts and explicit knowledge point constraints.

A structured domain-specific knowledge base was constructed to support generation. This knowledge base integrates past official examination papers, authoritative mock items, and associated knowledge point grids in a standardised and indexed format. During item generation, relevant entries are retrieved from the knowledge base according to the learner's weak knowledge point tags and task specifications, providing contextual grounding for the model's output.

Rather than relying on a static item bank, the system dynamically generates practice items in response to learner profile data retrieved from the back-end, while remaining aligned with examination standards through knowledge base grounding. This retrieval-based grounding mechanism ensures that generated items maintain structural validity and content alignment with the HSK Level 4 test blueprint.

The core logic of these models is dynamic adaptive generation. Upon receiving a personalised learner profile pushed from the back-end—containing elements such as a list of weak knowledge points, a set of mastered points, and the historical distribution of error types—the models prioritise generating items targeting the weak points. They simultaneously avoid simple repetition of consolidated knowledge, thereby enabling more targeted provisioning of practice content.

The Knowledge Point Recognition and Tagging Model, built on Doubao 1.5 Lite 32k, aligns its knowledge base strictly with the 2,500 core vocabulary items and 120 grammar points of the HSK Level 4 syllabus. Acting as the system's "data parser," its core function is to perform deep semantic analysis on answer results (particularly incorrect ones) uploaded from the front-end. The model identifies the specific test point and error type underlying a mistake. It outputs standardised data tags (e.g., {"status": "not mastered", "category": "grammar", "subtype": "concessive conjunction word order"}). These structured tags provide explicit decision-making criteria for the item generation models, translating vague "incorrect answers" into actionable intervention instructions. The key function of the middle-tier is to translate behavioural indicators into rule-governed instructional decisions.

3.2.3 Back-end layer: Dynamic data storage and relational modelling centre

The back-end utilises a relational database to construct a three-dimensional dynamic data model linking learners, knowledge points, and behaviours. Its design goes beyond static storage, aiming to achieve structured data integration, deep relational linking, and real-time iteration.

The database maintains three core categories of interrelated data entities. The Mastery Status Profile is continuously updated via a learner-knowledge point relational table, tracking each learner's status (i.e., mastered, needs consolidation, or not mastered) for all 2,500 vocabulary items and 120 grammar points, along with cumulative counts for times a student answers an item incorrectly and recent exposure timestamps, forming a dynamically evolving competency map. The Behavioural Sequence Log meticulously records a timestamped event stream of learning actions, including the start and end times of each practice session, time distribution across item types, and interaction depth with feedback pages (e.g., time spent viewing explanations). These data feed into models of learner engagement and study habits. Finally, the Error Tag Repository stores all error tags produced by the middle-tier recognition model, with each record linked to the specific item, knowledge point, and learner, creating a traceable history of error patterns. A core feature of this layer is its dynamic relational algorithm. The system establishes active links between different data entities based on predefined rule sets. For instance, a rule might automatically elevate a knowledge point to "high-priority weak point" status if its "not mastered" count exceeds 2 and the latest relevant response time exceeds 150% of the average for similar items. This transforms the database from a passive storage infrastructure into an active state-transition system, in which learner knowledge states evolve according to explicitly defined transition rules. It ensures that the intelligent agents in the middle-tier always access the most current and contextually relevant integrated data view of the learner, thereby providing a solid data foundation for the real-time optimisation of the learning path.

3.3 Core functions: The data-driven dynamic learning closed loop

All decision rules and weighting parameters were pre-defined prior to the formal intervention and remained constant throughout the experimental period. The core functions of the system revolve around a closed loop of data collection, analysis, intervention, and update. All functional designs are centred on data-driven automated learning, with the specific process as follows: Data-driven automated item generation for precise individualisation. Automated item generation is the core service of the system, and its logic is entirely data-driven, rather than relying on a static item bank: when a learner initiates a practice request, the middle-tier AI agent first calls the learner's latest data from the back-end database, including tags for weak knowledge points (e.g., "high-priority weak point - arranging words to form a sentence"), a list of mastered knowledge points, and the distribution of error types (e.g., "60% of errors are concentrated in grammar application").

Based on these data, the item generation models, in alignment with HSK Level 4 examination specifications and the "i+1" difficulty principle, generate targeted practice items. Knowledge points

marked as "not mastered" are assigned a higher generation weight (minimum 40%), while those classified as "to be consolidated" account for approximately 30% of the generated items. Knowledge points identified as "mastered" are sampled at a low rate (not exceeding 10%), with the remaining proportion allocated to adjacent or related knowledge points to support structured expansion. These proportional weights were determined through consultation with experienced HSK instructors and reflect established pedagogical practices in balancing reinforcement and progression. At the same time, they retain an exploratory character and were treated as fixed parameters during the formal intervention period to ensure procedural consistency.

To meet fragmented learning needs, the system analyses the "duration per practice session" data collected by the front-end (e.g., a historical average practice duration of 10 minutes) and automatically breaks down item sets into smaller chunks (5-8 items per set, taking 5-8 minutes). This ensures that practice is adapted to fragmented learning scenarios while maintaining systematic data recording and a targeted weekly exposure frequency.

Data-supported personalised feedback for precise error correction and attribution. The feedback function is designed around in-depth data analysis, rather than simply providing correct answers. Its design logic is "data tagging → feedback precision":

When a learner submits an answer, the system first compares the result with the standard answer to generate basic correctness data. It then triggers the knowledge point recognition model to perform a data analysis of the incorrect answer, generating error type tags (e.g., "lexical collocation error"), test point association tags (e.g., "related vocabulary: 承担 – 承受 "), and native language adaptation tags.

Based on these data tags, the system automatically generates layered feedback: a basic layer (correct answer + error type), an in-depth layer (test point analysis + example sentence extension), and an adaptive layer (native language translation and analysis). This ensures that the feedback content is tailored to the root cause of the learner's error and their level of understanding.

The feedback response time is optimised to be within 5 seconds through data coordination, ensuring that learners receive corrective information while their memory is still fresh. Meanwhile, the front-end collects data such as "feedback viewing duration" and "example sentence click counts" to subsequently optimise the adaptability of the feedback content.

Dynamic updates through data iteration for real-time learning path optimisation. Data iteration is the core guarantee for maintaining the system's personalisation, achieving a real-time closed loop of practice, data, and adjustment. After each practice session, the front-end automatically collects all data from that session (response results, time taken, feedback interaction behaviour) and synchronises it to the back-end database.

The middle-end intelligent agent triggers the data update mechanism. For correctly answered knowledge points, if answered correctly twice in a row, the status is updated from "to be consolidated" to "mastered", and the item generation weight is lowered. For incorrectly answered knowledge points, the number of times "not mastered" is recorded. If incorrect twice in a row, the knowledge point is upgraded to a "high-priority weak point", and the item generation weight is increased.

The back-end database updates the "learner–knowledge point" association table and the behaviour data model in real-time. This ensures that for the next automated item generation, the system can adjust the practice content and difficulty based on the latest data, avoiding repetitive practice on mastered knowledge points or omission of weak knowledge points, thus achieving dynamic optimisation of the learning path.

In summary, the core of this system's design is not an accumulation of digital tool functions, but the construction of a data-driven dynamic learning ecosystem. By collecting full-process learning data at the front-end, using data to achieve automated item generation and personalised feedback at the middle-end, and supporting decision-making through data modelling and real-time updates at the back-end, it

ultimately forms a closed loop of data collection, analysis, intervention, and update. This framework is intended to enable data to inform the entire learning process, thereby moving towards the automated and personalised learning goals of using data to guide practice allocation, locate weak areas, and optimise the learning path.

Importantly, the system does not assume that data replace pedagogy; rather, it encodes pedagogical principles into computable rules, allowing instructional theory to be enacted at scale.

4 Empirical Validation of the Data-Driven Intervention

To empirically validate the efficacy of the data-driven intelligent practice system in HSK Level 4 preparation, this study employed a quasi-experimental design centred on multi-dimensional data collection and analysis. A non-equivalent group pre-test-post-test framework was adopted, incorporating a natural usage-based grouping method to investigate the dose-response relationship of the data-driven intervention. Through strict control of extraneous variables and the use of standardised measurement tools, a complete empirical framework of data collection, difference comparison, and correlation validation was constructed to systematically examine the practical value of the data-driven pathway for language proficiency enhancement. The design follows a usage-based quasi-experimental paradigm frequently adopted in educational technology research, in which naturally occurring variation in system engagement is treated as a continuous treatment variable. The validation framework and results are as follows:

4.1 Participants and variable control

This study selected participants from parallel HSK Level 4 intensive classes at the Chinese Language Centre of a university in China, including a total of 49 international students. All participants were Russian native speakers of European nationality, aged 19-25 (M = 21.7, SD = 2.1). They had received 27 weeks of systematic Chinese instruction under a unified syllabus and were in the transition phase from HSK Level 3 to HSK Level 4. To ensure internal validity, key confounders were controlled: all three parallel classes shared the same instructor, schedules, teaching materials, and in-class activities.

All 49 participants were granted access to the intelligent practice system. Following the eight-week intervention period, learners were categorised into five tiers (Tier 1: ≤20 min; Tier 2: ~80 min; Tier 3: ~200 min; Tier 4: ~320 min; Tier 5: ≥400 min) based on their total back-end logged practice minutes per month. These tier thresholds were determined post-hoc by dividing the observed range of usage data into approximate quintiles, reflecting natural clusters of engagement behaviour rather than pre-assigned groups. Learners in "Tier 1: minimal users" were treated as a de facto control group. Due to minimal engagement, the system could not generate meaningful personalised interventions for this tier, making their learning experience a functional approximation of a low-intervention condition, as insufficient engagement limited the system's capacity to deliver sustained personalised adaptation. Learners in Tiers 2 to 5 constituted the experimental groups, receiving graded levels of data-driven intervention commensurate with their usage. To assess the initial equivalence of groups formed by this naturalistic process, a one-way ANOVA was conducted on the pre-test total scores across the five tiers. Although baseline proficiency equivalence was statistically supported, the possibility of self-selection bias inherent in voluntary engagement remains. Therefore, subsequent analyses focus on graded association patterns rather than binary group comparisons. The ANOVA result revealed no statistically significant difference across tiers (p = .48), supporting their comparability despite non-random assignment. Consequently, the primary analyses examined the dose-response relationship—specifically, the correlation between usage frequency and learning progress, as well as progress differences across dosage tiers—rather than relying

on simplistic experimental-versus-control mean comparisons. The potential moderating role of learner motivation in this relationship is theoretically discussed separately (see Section 5.4).

4.2 Data collection system

Proficiency Measurement Tools: Two standardised HSK Level 4 mock examinations were used as pre-test (mid-term exam) and post-test (final exam) tools. The test papers were compiled from genuine HSK Level 4 items and mock items of equivalent difficulty, calibrated using the Rasch model. To ensure the quality of the measurement tools, we assessed their reliability and validity. Internal consistency reliability, measured by Cronbach's alpha, was .87 for the pre-test and .89 for the post-test, indicating high reliability. Construct validity was supported by confirmatory factor analysis, which showed acceptable model fit indices for the predefined HSK Level 4 structure (e.g., CFI > .90, RMSEA < .08). Content validity was established through a review by three experienced HSK instructors who confirmed the items' coverage of the prescribed syllabus. The pre-test difficulty coefficient was b = -0.02, and the post-test difficulty coefficient was b = 0.01, with no statistically significant difference (t = .43, p = .67), ensuring the equivalence of the measurement tools. All raw scores were standardised to a maximum of 100 points to eliminate inconsistencies in scale due to differences in the papers, providing a uniform standard for data comparison.

Conceptualisation and Measurement of Intervention Intensity: The core independent variable is the intensity of the data-driven intervention, conceptualised as the degree to which a learner's practice is personalised and adapted based on the analysis of their own behavioural and error data. Directly measuring this multifaceted construct is complex. Therefore, we employed system usage frequency as a practical and quantifiable proxy indicator. Conceptually, usage frequency captures the behavioural exposure dimension of the intervention and serves as an observable proxy for the otherwise latent construct of adaptive personalisation intensity. The rationale is that higher usage generates more granular data, enabling the system to deliver more frequent and more precisely tailored interventions, i.e., a higher "dose" of data-driven intervention. Usage frequency was quantified using a 5-point Likert scale as mentioned above: Tier 1 (almost never use, ≤20 min/month), Tier 2 (occasional use, ~80 min/month), Tier 3 (average use, ~200 min/month), Tier 4 (frequent use, ~320 min/month), Tier 5 (daily use, ≥400 min/month). It should be noted that usage frequency primarily reflects cumulative exposure quantity and does not directly capture qualitative differences in engagement strategies.

Data Collection Dimensions: Three main categories of data were collected: 1) Proficiency improvement data (scores for each item type in the pre-test and post-test, total scores, and improvement amount ΔScore); 2) Intervention intensity data (system usage frequency level, actual practice duration); 3) Process data (distribution of error types for each item type, number of reinforcements for weak knowledge points), providing multi-dimensional support for subsequent correlation analysis.

4.3 Implementation of the data-driven intervention

The experimental procedure consisted of pre-test data collection, a data-driven intervention, and post-test data validation. Pre-test Phase: All learners took an HSK Level 4 mock examination to collect score data for the five major item types (cloze, sentence sequencing, reading comprehension, sentence formation, and picture-cued writing), which served as baseline proficiency data. Basic information about the learners was also recorded to establish an initial "learner–proficiency" data profile. Intervention Phase: Classroom instruction followed the regular schedule. All learners were provided access to the system for supplementary practice after class. Crucially, engagement was voluntary, simulating a natural learning environment. The system dynamically updated the learner's proficiency profile based on real-time practice data (response results, error types, practice duration) and used an automated item generation

algorithm to prioritise exercises corresponding to weak knowledge points. No additional incentives were provided for system use, and classroom assessment weighting did not depend on usage frequency, thereby reducing coercive engagement effects. Post-test Phase: All learners took a second HSK Level 4 mock examination. Learners were then grouped based on their back-end-logged usage frequency, forming the de facto control (Tier 1) and experimental groups (Tier 2-5). This usage data, combined with process data, formed a complete "intervention intensity–proficiency improvement" data matrix for analysis.

4.4 Validation of the data-driven effects

Of the 49 participants, 48 completed both the pre-test and post-test, yielding a valid sample rate of 97.9%. Data analysis focused on the correlation between the intensity (dose) of the data-driven intervention and proficiency improvement (response). The core results are as follows.

Descriptive statistics for the pre-test and post-test (see Table 1) show that the mean total score on the pre-test was 74.885 (SD = 18.043), which increased to 77.812 (SD = 12.965) on the post-test, a net improvement of 2.927 points. While modest in magnitude at the aggregate level, this increase should be interpreted in light of the substantial variability reduction observed. Furthermore, the dispersion of scores significantly decreased—the coefficient of variation dropped from 0.241 in the pre-test to 0.167 in the post-test. At the aggregate level, the pattern is compatible with both a modest overall score increase and a notable reduction in inter-individual variability, providing preliminary quantitative support for the hypothesised benefit of data-driven precise remediation. The reduction in score dispersion is consistent with the possibility that lower-proficiency learners benefited from targeted reinforcement, although individual-level gain patterns were not independently modelled.

Table 1

Descriptive Statistics for Pre-Test and Post-Test Scores

Variable	Max	Min	Mean	SD	Median	Kurtosis	Skewness	Coefficient of Variation
Pre-test	99.000	19.500	74.885	18.043	78.500	0.766	-1.016	0.241
Post-test	97.500	51.500	77.812	12.965	80.500	-0.909	-0.469	0.167

Note. Standardised scores were used, with a maximum score of 100.

Statistics on the improvement amount (ΔScore) for different usage frequency groups (see Table 2) reveal a significant positive association between the intensity of the data-driven intervention and learning progress, consistent with a dose-response relationship. Notably, the de facto control group (Tier 1, minimal intervention) showed a slight regression in total score ($\Delta = -3.768$). This contrasts with the gains observed in groups with higher system engagement and suggests that in the absence of sustained, targeted practice (whether data-driven or otherwise), knowledge attrition may occur, a phenomenon consistent with SLA theory. Tier 2 group (low-to-moderate dose) showed that overall progress turned positive, but the increase was limited (total score improvement of 1.125), indicating that low-frequency use can only maintain a basic learning state and may be insufficient to establish a stable data-driven remedial feedback loop. Tier 3 and above groups (moderate-to-high dose) showed significant growth. The Tier 5 group (highest dose) had the most outstanding progress, with a total score improvement of 30.5. These findings are compatible with the hypothesised mechanism of the intervention, in which increased data granularity enables more precise adaptation. Notably, the magnitude of gain in Tier 5 substantially exceeds that observed in adjacent tiers, suggesting a potential non-linear acceleration pattern at higher exposure levels.

Table 2

Descriptive Statistics of Score Gains for Various Item Types and Total

Tier	Cloze	Sentence Sequencing	Reading Comprehension	Sentence Formation	Picture-cued Writing	Total Score
Tier 1	-0.071	-0.357	-1.679	-0.536	0.054	-3.768
Tier 2	0.750	-1.000	0.750	0.750	0.500	1.125
Tier 3	2.300	2.600	1.900	1.400	1.350	11.550
Tier 4	2.250	4.000	3.000	3.750	0.625	16.250
Tier 5	5.000	6.000	3.500	8.000	4.500	30.500

Given the non-normal distribution of gain scores across tiers, a non-parametric Kruskal-Wallis test was conducted. The test revealed a statistically significant difference in total score improvement among the different usage frequency groups ($\chi^2 = 42.36$, $p < .001$). Post-hoc pairwise comparison with appropriate adjustment for multiple testing further indicated that the Tier 5 group differed significantly from the Tier 1-4 groups ($p < .01$), the Tier 4 group differed significantly from the Tier 1-2 groups ($p < .01$). These results suggest the existence of a potential "usage threshold" for observable progress: system usage of over 100 minutes per week (Tier 5) appears associated with significant progress.

Spearman's correlation analysis (see Table 3) revealed a significant positive correlation between system usage frequency and the improvement in scores for each item type, as well as the total score improvement ($p < .001$). The total score correlation coefficient reached 0.811. According to conventional effect size interpretations, this coefficient represents a strong association in educational research contexts. This strong, graded relationship provides convergent support for the structural alignment between intervention exposure and learning gain. The variation in correlation coefficients across item types provides a data-based rationale for future system optimisation, particularly for open-ended tasks.

Table 3

Spearman's Correlation between Usage Frequency and Scores on Various Item Types

	Cloze	Sentence Sequencing	Reading Comprehension	Sentence Formation	Picture-cued Writing	Total Score
Usage Frequency	0.641***	0.624***	0.774***	0.722***	0.46***	0.811***

Note. *** indicates statistical significance at the p < .001 level.

4.5 Summary of validation findings

The empirical data reveal a strong, graded positive association between the proxy for data-driven intervention intensity (system usage frequency) and learning outcomes. The dose-response pattern—with minimal gains at low usage and substantial gains at high usage, contrasted against the control group's stagnation—is compatible with a positive effect of the data-driven intervention. The high correlation ($r = .811$) indicates that system engagement is a strongly related factor to score improvement. However, given the quasi-experimental design with naturalistic grouping, a strict causal claim cannot be definitively made. Given the absence of baseline differences across tiers, the graded post-intervention divergence is unlikely to be attributable solely to pre-existing proficiency disparities. Future research employing randomised assignment or longitudinal cross-lagged modelling would be required to establish causal directionality. The observed relationship, while robust, may be partially confounded by unmeasured

variables such as intrinsic motivation or self-regulated learning ability. The findings therefore provide compelling evidence for a potent association and support the potential efficacy of the data-driven pathway within the constraints of this study. Importantly, the graded pattern observed across five dosage tiers reduces the likelihood that the findings are attributable to random fluctuation alone, particularly given the consistency across multiple item types and analytic approaches (group comparison and correlation analysis). The observed monotonic trend aligns with theoretical expectations of cumulative adaptive exposure. From a learning science perspective, the findings are theoretically coherent with cumulative practice models, which posit that adaptive reinforcement effects compound across repeated exposure cycles.

5 Discussion and Conclusion

5.1 Data-driven adaptation as mechanism for addressing learner heterogeneity

The findings reveal a robust graded association between system usage intensity and HSK Level 4 score gains. Learners in the highest usage tier achieved an average increase of 30.5 points, 27 times that of the lowest-frequency group, demonstrating a pattern analogous to the "dose-response" effect reported in adaptive learning research (Ma et al., 2014; Zawacki-Richter et al., 2019). Rather than reflecting mere exposure time, this gradient suggests cumulative advantages derived from sustained participation in feedback-integrated adaptive cycles.

From a theoretical perspective, the results illustrate how data-driven systems operationalise individual adaptation. By continuously aggregating behavioural traces, error distributions, and proficiency indicators, the system constructs a dynamically updated learner model that informs subsequent task allocation. This aligns with learning analytics scholarship emphasising the transformation of trace data into actionable pedagogical decisions (Ferguson, 2012). The closed-loop structure, characterised by diagnosis, targeted reinforcement, and iterative recalibration, reflects adaptive feedback models central to formative assessment theory (Hattie & Timperley, 2007; Panadero & Lipnevich, 2022).

The concentration of 60% of "sentence formation" errors within "word order logic" exemplifies this mechanism. Algorithmic reweighting of this knowledge point resulted in an 8-point gain among high-frequency users, substantially exceeding improvements observed in traditional settings. Such outcomes reflect algorithm-supported mastery progression, whereby targeted corrective practice incrementally reduces specific knowledge gaps while preserving overall efficiency (Holmes et al., 2019). The improvement observed, therefore, is structural rather than incidental: it derives from the optimisation of adaptive feedback loops rather than increased practice volume alone.

5.2 Reconfiguring Pedagogical Authority: From Experience-Driven to Evidence-Based Instruction

Conventional Chinese language instruction has largely operated within an experience-driven paradigm, where pedagogical decisions are grounded in teacher intuition. While professional expertise remains valuable, reliance on subjective judgment may obscure latent learner heterogeneity (Ifenthaler & Yau, 2020). In the present context, the traditional classroom demonstrated an average weekly recurrence of 2.7 exposures for core vocabulary, which falls below consolidation thresholds commonly cited in second language acquisition research, yet this limitation remained undetected without systematic data monitoring.

With 49 learners exhibiting substantial baseline dispersion (pre-test coefficient of variation = 0.241), uniform practice structures inevitably produced redundant exposure for advanced learners and insufficient reinforcement for those struggling. Learning analytics studies similarly note that aggregate

instruction can mask individual trajectories, thereby constraining differentiated scaffolding (Viberg et al., 2018; Matcha et al., 2019).

The data-driven model, by contrast, redistributes decision authority toward learner-generated evidence. Continuous modelling of 2,500 lexical items and 120 grammatical points enabled adaptive adjustment of recurrence frequency to an average of 7.3 exposures for high-priority weaknesses. Conceptually, this reflects an operationalisation of mastery learning principles (Bloom, 1968; Guskey, 1980), implemented at micro-level granularity through algorithmic modelling.

Importantly, adaptation reduced not only mean performance deficits but also achievement dispersion (post-test coefficient of variation = 0.167). This contraction of variance suggests that data-driven intervention may function as a gap-moderating mechanism, consistent with meta-analytic evidence indicating that targeted feedback disproportionately benefits lower-performing learners (Panadero & Lipnevich, 2022). The shift from experience-driven to data-driven instruction thus represents a structural reconfiguration of pedagogical authority, moving from intuition-led uniformity to evidence-based differentiation.

5.3 Beyond digitalisation: datafication as structural transformation

Digital transformation in International Chinese Language Education is frequently conflated with the mere digitalisation of instructional materials or the migration of content to online platforms. However, digital delivery alone does not constitute pedagogical innovation (Bond et al., 2020). Platforms that replicate paper-based exercises in digital form without analytics-driven optimisation rarely alter underlying instructional logic.

The present findings underscore the distinction between digitalisation and datafication. Datafication entails reconstructing instructional processes around continuous data generation, modelling, and intervention. Rather than serving descriptive or archival purposes, behavioural data become prescriptive inputs for algorithmic decision-making (Ifenthaler & Yau, 2020; Matcha et al., 2019). In this system, front-end interactions are transformed into automated item allocation, personalised feedback, and dynamic pathway recalibration.

The modularisation of exercises into 5-8 item segments, derived from practice duration analytics, enabled over 90% of learners to complete effective micro-practice sessions. This design illustrates how behavioural modelling can sustain engagement in fragmented learning environments. The transition from digitalisation to datafication, therefore, constitutes a structural shift in instructional logic: data evolve from passive records into active drivers of adaptive progression.

5.4 Boundary conditions: Motivation as a moderating mechanism

Despite the strong correlation between usage frequency and performance gains (r = .811), behavioural metrics do not inherently establish causality. Engagement indicators may reflect pre-existing motivational and self-regulatory differences (Viberg et al., 2018; Ifenthaler & Yau, 2020).

Learners in higher tiers demonstrated strategic engagement by analysing feedback, adjusting difficulty levels, and allocating extended time to error review, thereby activating a self-regulated feedback loop. In contrast, minimal users engaged superficially and repeated 41% of identical knowledge-point errors, indicating limited metacognitive monitoring. This divergence aligns with research identifying feedback utilisation and self-regulation as mediators between technological affordances and learning outcomes (Panadero, 2017).

The regression observed among the lowest-tier learners likely reflects the interaction between low foundational proficiency and insufficient motivational engagement rather than technological inefficacy. These findings suggest a boundary condition: data-driven systems optimise pathways but do not

substitute learner agency. Technology amplifies existing motivation; it does not generate it. Effective implementation must therefore integrate adaptive analytics with pedagogical strategies that cultivate sustained engagement and strategic learning behaviours.

5.5 Limitations, implications and future directions

Several limitations warrant consideration. First, the sample consisted exclusively of Russian native-speaking learners within a single institutional context, constraining external validity. Cross-linguistic replication is necessary to examine whether similar adaptation patterns emerge across diverse learner populations.

Second, outcome measures focused primarily on short-term score gains. The absence of delayed post-tests limits inferences regarding long-term retention and transfer to communicative competence in authentic contexts.

Third, the system's modelling capacity remains more robust for structured items than for open-ended productive tasks. Semantic appropriateness and pragmatic felicity in discourse-level production require more advanced natural language processing frameworks than those currently implemented.

This study contributes to the convergence of adaptive learning, mastery learning, and learning analytics within second language education. By demonstrating a graded relationship between intervention intensity and learning outcomes, as well as a reduction in performance variance, it provides empirical grounding for precision-oriented instructional models in high-stakes proficiency preparation.

Future research should extend participant diversity, incorporate longitudinal designs to examine retention trajectories, and refine modelling approaches for evaluating semantic and pragmatic dimensions of open-ended production. Additionally, investigating hybrid frameworks that integrate teacher expertise, motivational scaffolding, and algorithmic precision may illuminate how human and technological agency can be optimally aligned.

6 Conclusion

Focusing on HSK Level 4 preparation, this study examined the instructional efficacy of a data-driven adaptive practice system. The findings demonstrate a graded association between intervention intensity and learning gains, accompanied by reduced performance dispersion among learners.

More broadly, the results suggest that datafication represents a structural transformation in instructional logic, shifting pedagogical decision-making toward learner-generated evidence and adaptive modelling. However, technological optimisation alone is insufficient; its effectiveness is contingent upon learner motivation and strategic engagement.

By clarifying both the potential and the boundary conditions of data-driven adaptation, this study advances the conceptualisation of precision-oriented language education in digitally mediated environments.

References

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*(2), 1–12.

Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., & Kerres, M. (2020). Mapping research in student engagement and educational technology in higher education: A systematic review. *International Journal of Educational Technology in Higher Education, 17*, 2. https://doi.org/10.1186/s41239-019-0176-8

CCTV.com. (2025). 全球说国际中文教育智能体在京发布 [Global Talk international Chinese language education intelligent agent released in Beijing]. http://www.cctvzw.org.cn/hydt/1602.html

Eyal, L. (2012). Digital Assessment Literacy — the Core Role of the Teacher in a Digital Environment. *Journal of Educational Technology & Society*, *15*(2), 37–49. http://www.jstor.org/stable/jeductechsoci.15.2.37

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5–6), 304–317. https://doi.org/10.1504/IJTEL.2012.051816

Guskey, T. R. (1980). Mastery learning: Applying the theory. *Theory Into Practice, 19*(2), 104-111.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education promises and implications for teaching and learning.* Center for Curriculum Redesign.

Hu, Jiehui (胡杰辉), & Zhang, Tiefu (张铁夫). (2023). 中国高校外语教师数字素养的信念与实践研究 [A study on the beliefs and practices of digital literacy among foreign language teachers in Chinese universities]. 外语与外语教学[Foreign Language and Foreign Language Teaching], *44*(5), 73-85.

Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development, 68*, 1961–1990.

Krashen, S. D. (1982). *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon.

Krashen, S. D. (1985). *The Input Hypothesis: Issues and Implications.* London: Longman.

Li, Quan (李泉). (2020). 国际中文教育转型之元年 [The inaugural year of the transformation of international Chinese language education]. 海外华文教育 [Overseas Chinese Education], (3), 3-10.

Liang, J. C., Hwang, G. J., Chen, M. R. A., & Darmawansah, D. (2021). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2021.1958348

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In Ritchie & Bhatia (Eds.), *Handbook of Second Language Acquisition* (pp. 413-468). Academic Press.

Ma, W., Adesope, O., Nesbit, J., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918. https://doi.org/10.1037/a0037123

Matcha, W., Gašević, D., Uzir, N. A. A., Jovanović, J., & Pardo, A. (2019, March). Analytics of learning strategies: Associations with academic performance and feedback. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 461-470).

Meng, Y., & Zhang, C. (2024). Move Up the Fluency Ladder: Introducing Deep English as an Artificial Intelligence Speaking Tool for Speakers of English as a Foreign Language. RELC Journal. https://doi.org/10.1177/00336882241252496

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. F*rontiers in psychology, 8*, 422.

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review, 35*, 100416. https://doi.org/10.1016/j.edurev.2021.100416

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*(2), 72-78. https://doi.org/https://doi.org/10.1016/0346-251X(78)90027-1

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129-158.

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior, 89*, 98–110.

Wang, Hanwei (王汉卫), & Zhang, Xinyue (张馨月). (2024). 论"人群特征"对国际中文教育的学科支撑 [On the disciplinary support of 'population characteristics' for international Chinese language education]. 语言战略研究[Language Strategy Studies], *9*(5), 54-63.

Xinhua News Agency. (新华社). (2024). 大数据背景下大学英语教育数字化转型 [Digital transformation of college English education in the context of big data]. https://www.news.cn/expo/20240902/6387a8ed40b84dbe8553f6a8a879338b/c.html

Yuan, Xi (袁羲), & Wu, Yinghui (吴应辉). (2023). ChatGPT Plus 给国际中文教育带来的机遇、风险及应对策略 [Opportunities, risks, and coping strategies brought by ChatGPT Plus to international Chinese language education]. 云南师范大学学报（国际中文教育教学与研究版）[Journal of Yunnan Normal University (International Chinese Language Education and Research Edition)], (3), 53-62. https://doi.org/10.16802/j.cnki.ynsddw.2023.03.011.

Zawacki-Richter, O., Marín, V.I., Bond, M. et al. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators?. *International Journal of Education Technology in Higher Education, 16*, 39. https://doi.org/10.1186/s41239-019-0171-0

Zhang, C., Meng, Y., & Ma, X. (2024). Artificial intelligence in EFL speaking: Impact on enjoyment, anxiety, and willingness to communicate. *System* 121: Article 103259. https://doi.org/10.1016/j.system.2024.103259

Nanxi Bian is a PhD candidate at the Department of English at University of Macau. Her research interests include language education, language assessment literacy, and theoretical linguistics.

Qingyu Gao, PhD, is a Lecturer at the Chinese Language Centre of Shenzhen MSU–BIT University. His research focuses on discourse studies, second language acquisition, artificial intelligence in education, adaptive learning systems, and learning analytics. He is particularly interested in data-informed instructional design and the development of intelligent systems for international Chinese language education.

国际汉语教学学报（2026）
第 07 卷 · 第 03 期 133-152 https://doi.org/10.46451/ijclt.260411

数据驱动的语言能力发展路径构建
——基于 HSK4 级 AI 自动出题工具的干预研究

边楠茜
澳门大学，中国澳门特别行政区

高清宇
深圳北理莫斯科大学，中国

摘要
传统的 HSK 备考模式普遍采用同质化的训练方式，难以充分回应学习者个体差异，进而可能制约学习者语言能力的发展。本研究考察一项面向 HSK4 级的数据驱动学习干预的有效性。该干预依托自动出题系统实施，在二语习得理论的指导下，借助大数据分析诊断学习者的错误类型与薄弱环节，并针对 HSK4 级核心题型动态生成个性化 “i+1” 练习，以实现精准化学习干预。研究以 49 名俄语母语来华留学生为对象，开展了为期 8 周的准实验研究，并依据自然使用强度进行分组，考察系统参与度与学习成效之间的关联。量化结果表明，系统使用强度与成绩提升呈显著正相关（$r = .811$，$p < .001$）。研究结果表明，数据驱动干预可为缓解同质化教学的问题提供智能化解决方案。通过实现对学习过程的实时追踪、诊断与个性化反馈，本文提出了一条数据驱动的语言能力发展路径，可为 HSK 备考实践及语言教育数字化转型提供参考。

关键词
数据驱动语言学习，HSK4 级，智能教学系统，二语习得，个性化学习

边楠茜，澳门大学人文学院英文系在读博士生，主要研究方向为二语习得、技术赋能外语教学、二语写作。

高清宇，博士，现任深圳北理莫斯科大学汉语中心讲师。其研究方向涵盖话语研究、二语习得、教育人工智能、自适应学习系统与学习分析，尤其关注数据驱动的教学设计及国际中文教育智能系统的开发。