International Journal of Assessment Tools in Education
2026, Vol. 13, No. 1, 330–358
https://doi.org/10.21449/ijate.1726804
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Developing and validating the Second Language Buoyancy Scale (L2BS): Evidence from psychometric and machine learning analyses

Kenan Gao[1,2], Juan Zhang[1,2*], Yihui Wang[1,2], Wei He[3], Jianhong Mo[4]

[1]Faculty of Education, University of Macau, Macau, China
[2]Centre for Cognitive and Brain Sciences, University of Macau, Macau, China
[3]Department of Engineering Science, Faculty of Innovation Engineering, Macau University of Science and Technology, Taipa, Macau, China
[4]Department of Psychology, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE HISTORY

Received: June 25, 2025

Accepted: Dec. 2, 2025

KEYWORDS

L2 buoyancy,
Academic buoyancy,
Second language acquisition,
Scale development,
Machine learning.

Abstract: Given that existing academic buoyancy measures do not capture learners' everyday capacity to cope with setbacks in the L2 learning, an L2-specific scale is needed to assess second language (L2) buoyancy. This study aimed to develop and validate the Second Language Buoyancy Scale (L2BS). Using convenience sampling, data were collected from 554 university students at two mainland Chinese institutions and randomly split into two equal subsets (n = 277 per subset). Content validity was established via qualitative item generation (17 interviews) and expert review (ICC = .83). For structural validity, EFA on Subset 1 (KMO = .826; Bartlett's $\chi^2(6) = 616.99$, $p < .001$) supported a single-factor, four-item solution with loadings > .65; CFA on Subset 2 showed good fit (RMSEA = .096, CFI = .992, TLI = .977). Internal consistency was strong (Cronbach's α = .898; McDonald's ω = .898). Construct validity was supported by AVE = .689 and small-to-moderate correlations with academic buoyancy, growth mindset, grit, and conscientiousness. Criterion-related validity was evidenced by hierarchical regressions (incremental variance: $\Delta R^2 = .173$ for L2 engagement; $\Delta R^2 = .160$ for L2 enjoyment) and machine-learning models (Random Forest/XGBoost/LightGBM), in which L2BS consistently outperformed academic buoyancy (best accuracies: 73.21% for engagement; 64.29% for enjoyment). Overall, L2BS provides a brief, reliable, and valid measure of L2 buoyancy with clear utility for predicting key L2 outcomes such as L2 engagement and L2 enjoyment.

INTRODUCTION

Socioemotional factors play an important role in shaping academic performance (Farrington *et al.*, 2012). One such socioemotional learner-specific factor gaining increasing attention in second language acquisition (SLA) research is academic buoyancy (AB) (Sudina & Plonsky, 2021). AB refers to "students' capability to deal with challenge, adversity, and setback" (Martin & Marsh, 2006, p. 268), which are "typical of the ordinary course of school lif" (Martin & Marsh, 2008a, p. 54) or "everyday resilience" (Martin & Marsh, 2008b, p. 168) such as anxiety during exams or upset in a grade (Martin & Marsh, 2009). This aligns academic buoyancy with

*Corresponding author at: Juan ZHANG juanzhang@um.edu.mo University of Macau, Faculty of Education, Macau, China

e-ISSN: 2148-7456

the principles of positive psychology, which emphasizes individual strengths, well-being, and resilience (Martin & Marsh, 2008a).

While educational studies widely acknowledge the significant contribution of AB (Collie *et al.*, 2017) to enhance learning efficiency and progress across various disciplines and age groups, its role in foreign or second language (henceforth, L2) learning environments remains underexplored and underappreciated. Typically, researchers use the concept of AB in L2 learning studies without treating L2 buoyancy as an independent construct. According to Teimouri *et al.* (2020, p. 895), "the limitation of past personality research in SLA lies in its application of global personality assessment tools used in social psychology." Language-domain-specific measures of personality could offer a more precise assessment of students' personality traits and behavioral features in L2 settings than out-of-context, domain-general personality measures (Teimouri *et al.*, 2020). Additionally, there have been domain-specific language-psychological traits in the practice, for example, L2 grit (Sudina & Plonsky, 2021). AB can also be used in conjunction with other positive concepts (e.g., intended effort and grit) as important noncognitive individual traits in L2 learning (Sudina & Plonsky, 2021), as the ability of L2 learners to successfully master a language largely depends on their capacity to manage challenges, adversity, and setbacks.

Therefore, the present study aimed to: (1) develop a domain-specific Second Language Buoyancy Scale (L2BS) and examine its content, structural, and construct validity (convergent, discriminant, and criterion-related), as well as its reliability; (2) examine its criterion-related and predictive validity for key L2 outcomes, including via both traditional hierarchical regression and machine learning–based predictive models.

Academic Buoyancy

AB has emerged as a focal point of interest within educational psychology. AB is distinct from, yet complementary to, the broader construct of academic resilience (Martin & Marsh, 2009). The differences manifest in three primary ways: (1) Extent. AB specifically addresses everyday challenges, whereas academic resilience concerns significantly more severe adversities that pose major disruptions to educational progress, such as serious personal crises or extended school absences (Martin *et al.*, 2010; Martin & Marsh, 2009). (2) Frequency. AB is concerned with the regular and ongoing challenges that all students might face as part of their daily educational experience, and academic resilience deals with rarer, high-impact events (Martin *et al.*, 2010; Martin & Marsh, 2009). (3) Scope. Academic resilience refers to overcoming acute and chronic adversity impacting the educational domains (Martin *et al.*, 2010), with influence on a smaller segment of students compared to AB. In other words, AB has a broader application than academic resilience (Rudd *et al.*, 2021).

AB has been recognized as a multidimensional construct. While Martin and Marsh (2006) proposed the well-known 5C model - control, confidence, coordination, commitment, and composure - other researchers have emphasized related constructs such as self-efficacy (Lei *et al.*, 2022), teacher-student relationships (Hejazi & Abbasi, 2021), and strength-based parenting (Gu *et al.*, 2023), reflecting both personal and contextual influences on AB.

Situated within control-value theory (Pekrun, 2006) and self-determination theory (Ryan & Deci, 2024), AB operates by sustaining perceived control and value, thereby enabling adaptive emotion regulation and metacognitive responding following minor failures. Through resource-accumulation processes based on broaden-and-build theory (Fredrickson, 2001), AB helps preserve positive affect and task-focused behavior, thereby linking everyday challenges to downstream academic and psychological outcomes. This account also clarifies boundary conditions, such as classroom evaluation climate (Hoferichter *et al.*, 2021) and teacher support (Granziera *et al.*, 2022), that moderate the strength of AB's effects.

Therefore, in L2 learning, when students feel that L2 tasks are manageable and worthwhile, they are more likely to experience enjoyment and hope rather than anxiety or hopelessness. From a Self-Determination Theory perspective (Ryan & Deci, 2024), sustained motivation depends on the satisfaction of basic psychological needs for autonomy, competence, and relatedness. L2 buoyancy can thus be conceptualized as a capacity to maintain adaptive control-value appraisals and to protect need satisfaction in the face of everyday L2 setbacks. Buoyant L2 learners are better able to regulate negative emotions, preserve a sense of competence after errors, and remain autonomously engaged with L2 tasks despite temporary failures.

Empirical studies have employed diverse methodologies, including longitudinal (e.g., Putwain *et al.*, 2015; Hirvonen *et al.*, 2020), cross-sectional (e.g., Collie *et al.*, 2015; Datu & Yang, 2018), and experimental designs (e.g., Ren *et al.*, 2025). In terms of population, most prior studies have targeted adolescent and undergraduate student samples across diverse cultural contexts, such as Australia, the UK, Finland, and Asian countries. The current empirical results have consistently demonstrated positive associations between AB and various academic and psychological outcomes. Specifically, higher levels of AB have been linked to better academic performance (Collie *et al.*, 2015; Putwain & Daly, 2013), reduced academic stress and anxiety (Hirvonen *et al.*, 2019; Putwain *et al.*, 2015), lower boredom and hopelessness (Hirvonen *et al.*, 2020), greater learning enjoyment (Hirvonen *et al.*, 2020), higher motivation and engagement (Datu & Yang, 2018), and higher grit levels (Bayrami *et al.*, 2012). Importantly, many of these outcomes not only correlate with AB but may also act as antecedents or mediators, suggesting complex reciprocal relationships.

Collectively, this body of research underscores that AB is not merely a function of initial motivation but reflects a sustained capacity to regulate motivation and adaptively manage everyday academic demands (Yang *et al.*, 2024).

L2 Buoyancy

The importance of L2 buoyancy was first recognized by Yun *et al.* (2018). L2 represents a distinct dimension of L2 motivation, capturing learners' capacity to effectively manage everyday academic setbacks and challenges in SLA. This construct, conceptually situated at the intersection of positive psychology and SLA, emphasizes learners' proactive emotional regulation and sustained motivation beyond mere linguistic proficiency. Empirical evidence indicates that L2 buoyancy predicts both domain-specific (e.g., TEPS exam scores) and general academic achievement (e.g., GPA) (Yun *et al.*, 2018), while also contributing to psychological strengths such as L2 grit (Yang *et al.*, 2024).

Despite the growing recognition of its importance, measurement of L2 buoyancy remains limited. Most studies have adapted the Academic Buoyancy Scale (ABS) originally developed by Martin and Marsh (2008a) for general educational contexts. For instance, Fu (2024), Yun *et al.* (2018), and Jahedizadeh *et al.* (2019) applied modified versions of the ABS for L2 learners, often with minimal adaptation. While these studies reported acceptable reliability and some forms of validity (e.g., internal consistency, part of construct validity), critiques have emerged regarding the psychometric rigor of such adaptations (Yang *et al.*, 2024; Sudina & Plonsky, 2021).

Specifically, existing adaptations often rely on limited item modifications without a comprehensive evaluation of content validity, dimensional structure, or measurement invariance across different L2 learner populations. For example, Sudina and Plonsky (2021) extended Martin and Marsh's (2008a) original four-item scale by simply adding one additional dimension and four new items, without providing systematic theoretical justification or conducting robust factor analytic procedures. Thus, while these adapted scales demonstrate some degree of reliability and face validity, they fall short of meeting advanced psychometric standards, such as thorough construct validation (convergent, discriminant, and factorial validity), test-retest reliability, and cross-sample generalizability.

L2 buoyancy shares the same higher-order mechanisms with AB (Pekrun, 2006; Ryan & Deci, 2024) but is domain-specific to language learning. Frequent public corrective feedback, face threats, and situational uncertainty amplify the emotional load of micro-setbacks (Wiboolyasarin *et al.*, 2022). Accordingly, L2 buoyancy hinges on fluctuations in perceived control and value during communicative tasks and is jointly shaped by error-correction style (Wiboolyasarin *et al.*, 2022), teacher-student relations, and peer safety (Luo & Derakhshan, 2024).

In contrast to general academic settings, L2 learning is characterized by frequent public corrective feedback, face-threatening situations, and uncertainty in real-time communication. These context-specific stressors heighten learners' need to regulate their emotions, maintain perceived control over L2 tasks, and sustain motivational persistence in the face of repeated micro-failures. Consequently, a domain-general buoyancy construct may underestimate the intensity and dynamics of buoyant responding in L2 classrooms. A language-specific buoyancy construct is therefore needed to capture how learners cope with the recurrent, socially embedded challenges unique to L2 learning.

The absence of a psychometrically robust, L2-specific buoyancy instrument has at least three consequences for SLA research. First, it precludes precise estimation of the unique contribution of L2 buoyancy to language-specific outcomes such as engagement, enjoyment, and performance, over and above general academic buoyancy and related traits. Second, it limits researchers' ability to test theoretically grounded process models that link classroom practices (e.g., feedback style, teacher support) to L2 emotions and motivation via buoyant coping. Third, it constrains the design and evaluation of interventions aimed at strengthening learners' everyday resilience in L2 classrooms, because changes in L2 buoyancy cannot be reliably detected. Addressing this measurement gap is therefore essential for advancing both theory and practice in positive L2 psychology.

Research Model

A nomological framework (Figure 1) was presented to motivate the planned analyses of the construct, in which L2 buoyancy is positively associated with other adaptive dispositions (e.g., grit, growth-oriented language learning mindset, conscientiousness, and general academic buoyancy) and positively predicts L2 engagement and enjoyment (Figure 1). These hypothesized relations serve both to locate L2 buoyancy within the broader landscape of positive L2 psychology and to provide multiple sources of construct and criterion-related validity evidence for the newly developed scale.

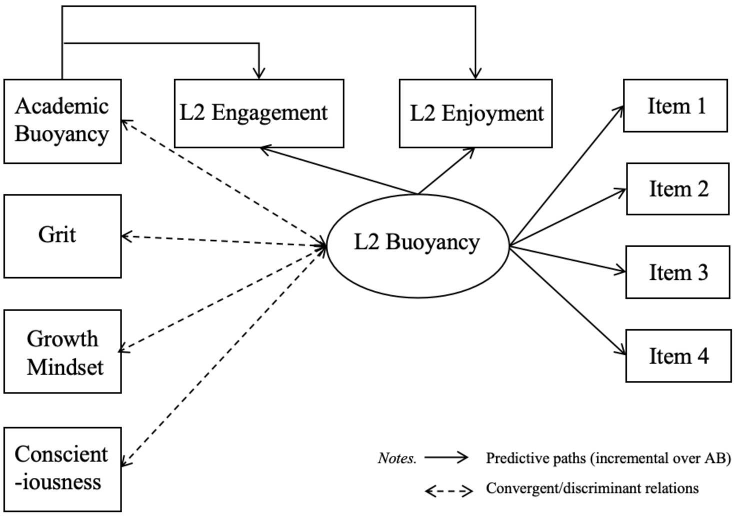


Figure 1. *Measurement model and validity framework for L2 Buoyancy.*

The Present Study

Recognizing these limitations, the present study seeks to develop and validate a psychometrically robust L2BS among Chinese undergraduate EFL learners from the ground up. The methodology adhered to established protocols for scale creation outlined by Clark and Watson (2016), ensuring a comprehensive evaluation of the scale's validity in multiple dimensions. The conceptualization of L2 buoyancy was grounded in the positive psychology framework (Seligman & Csikszentmihalyi, 2014) and previous models of AB (Martin & Marsh, 2008a), which highlight learners' capacity to adaptively manage everyday academic challenges in the SLA context.

In developing the scale, substantive validity was first addressed through interviews and expert evaluation to define the construct and generate an initial item pool reflecting key dimensions identified in prior research (Martin & Marsh, 2006). To establish structural validity, we employed both Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). EFA (using Principal Components Analysis [PCA] for data reduction and initial structure identification) was conducted to explore the underlying factor structure of the newly developed scale, given the lack of prior empirical consensus regarding the dimensionality of L2 buoyancy. PCA was selected in the early stage as it helps identify maximal variance components, serving as a preliminary step before applying stricter common factor models in CFA. Following EFA, CFA was used to test hypothesized factor structures derived from both theoretical expectations and EFA findings, with model fit evaluated through established indices (e.g., Comparative Fit Index [CFI], RMSEA, SRMR). Reliability was assessed through internal consistency coefficients (Cronbach's alpha), while multiple forms of construct validity were examined. Convergent validity was evaluated through Average Variance Extracted (AVE), ensuring that items within each factor captured shared variance. Discriminant validity was assessed by comparing inter-factor correlations and AVE scores. Criterion-related validity was explored by correlating L2BS scores with related constructs, including grit, language learning mindset, conscientiousness, and academic buoyancy, based on established links in SLA research (Meyer *et al.*, 2019; Teimouri *et al.*, 2020; Vrdoljak *et al.*, 2018; Yang *et al.*, 2024).

Importantly, this study also integrated machine learning (ML) approaches to examine the predictive validity of the L2BS by comparing its capacity to predict L2 engagement and enjoyment against the existing ABS. ML excels in processing large datasets, enhances prediction accuracy (Rezaii *et al.*, 2019). ML is increasingly adopted in psychology and applied linguistics to enhance measurement accuracy, efficiency, and out-of-sample prediction for psychometric and language-learning assessments (Lin, 2021; Shatte *et al.*, 2019). Flexible algorithms (e.g., random forests) recover nonlinear effects and higher-order interactions that conventional parametric models may miss (Shatte *et al*., 2019). Unlike linear regression, which is constrained by assumptions of linearity and normality, ML offers assumption-light function approximation and a prediction-first framework that can refine theory through accurate forecasting (Yarkoni & Westfall, 2017). Following Gao *et al.* (2025), we apply ML to the L2BS to uncover L2 learning, relevant predictors overlooked by traditional methods, and to improve predictive precision. This innovative step enabled a more nuanced evaluation of the practical utility of the L2BS in real-world language-learning contexts.

This initiative not only broadened the conceptual understanding of L2 buoyancy but also marked an effort to evaluate this construct quantitatively. Diverging from previous research, this study innovatively integrated machine learning techniques to examine the differential predictive power of ABS and L2BS in fostering L2 engagement and enjoyment, thereby establishing the scale's relevance in language-learning settings. By addressing existing measurement gaps, this research aims to contribute a theoretically grounded, empirically validated instrument for future SLA research and practice.

The following research questions were proposed:

RQ1: What is the underlying factor structure of the L2BS among Chinese undergraduate EFL learners?

RQ2: To what extent does the L2BS demonstrate adequate reliability (e.g., internal consistency)?

RQ3: Does the L2BS show evidence of convergent and discriminant validity in relation to conceptually related constructs such as grit, language learning mindset, conscientiousness, and general academic buoyancy?

RQ4: Does the L2BS demonstrate criterion-related and predictive validity for L2 engagement and enjoyment, and does it provide incremental predictive power over and above general academic buoyancy in hierarchical regression analyses?

RQ5: How does the predictive performance of the L2BS compare with that of the Academic Buoyancy Scale (ABS) when evaluated using machine learning models (Random Forest, XGBoost, LightGBM)?

PHASE 1: ITEM POOL OF L2BS

Method

Participants

Based on the concept of L2BS, participants were purposefully selected from college students who had faced difficulties but shown resilience in learning a foreign language. In coordination with instructors from two universities in mainland China, 29 English-as-a-second-language students were initially identified. After conducting 12 interviews, the researchers determined that data saturation had been reached. To ensure the data's robustness, five additional interviews were carried out, bringing the total to 17. The final sample consisted of 15 undergraduate and 2 master's students (2 males and 15 females), with an average age of 19.47 years (SD = 1.12).

Interview procedure

All the interviews were conducted online by four research assistants via Tencent Meeting. Prior to the interviews, all participants provided informed consent, acknowledging the academic nature of the study and its independence from their educational institutions. All participants consented to the audio recording of the interviews. To ensure consistency across interviews, a structured protocol with six questions was employed (refer to Appendix 1), with two core inquiries centering on participants' retrospective accounts of daily obstacles encountered in their English language learning journey from childhood to adulthood, their strategies for overcoming these challenges, and their associated emotions. The interviews were conducted in Mandarin and averaged approximately 20 minutes in duration.

Analysis and results

The interview analysis began with a systematic process for handling the original audio recordings. (1) Transcription: Each interview was promptly transcribed into text to maintain the accuracy and integrity of the data. (2) Content Analysis: The transcript text was analyzed using a step-by-step content-coding. Text materials were first categorized based on content. Ambiguous or unclear information was independently coded and then reviewed for a secondary step, incorporating it into the initial categories after determining its meaning. Following these steps, the data was organized into only one category. To ascertain the content validity, the one category comprising four items underwent evaluation by feedback from interviewees.

In accordance with the guidelines proposed by Marsh *et al.* (2005), items within the same factor must meet specific criteria. The criteria for item selection can be summarized as follows: Firstly, items should display significant item-total correlations. Secondly, they should demonstrate robust factor loadings on the designated factor of interest while showing minimal cross-loadings on other factors. Lastly, selected items should closely align with the theoretical framework of the focal factor and the broader construct. To validate the initial items based on these established criteria, each factor and its constituent items underwent a thorough examination using SPSS 26.

A panel consisting of five experts (two professors in Applied Linguistics, two senior PhD candidates in Language Education, and one senior PhD candidate in Educational Psychology) was asked to assess the items to ensure their content validity. The expert assessment (Appendix 2 lists the rating rubric) involved employing intraclass correlation coefficients (ICCs) alongside two-way mixed models, and the mean of the five raters was computed to gauge absolute inter-rater agreement. Following the expert panel's evaluation, it was concluded that no items required deletion, and the ICC for the 4-item scale was determined to be .83. Consequently, the initial assessment of L2 buoyance utilized a 7-point Likert scale scoring system, with ratings ranging from 1 (Strongly disagree) to 7 (Strongly agree). Appendix 3 lists the specific items of the initial L2BS.

PHASE 2: INITIAL TEST OF L2BS

Method

Participants

Data were collected via online questionnaires using convenience sampling from 554 participants across two mainland Chinese universities. All participants had taken the Gao Kao (China's university entrance exam, which includes an English exam) and began learning English in primary school, continuing with compulsory English courses at the university level. The sample primarily consisted of undergraduate students (N = 531, 95.85%), aged 18-25 (N = 546; Male = 144, Female = 410), with 14 graduate and 9 doctoral students. Participants represented diverse disciplines, including humanities (4.33%), social sciences (13.72%), natural sciences (0.7%), engineering (1.99%), and medical (79.24%).

The initial dataset is divided randomly into two equivalent parts (277 each). A t-test was conducted to compare the demographic characteristics (e.g., age, major) and psychological and personality traits between the two groups, ensuring no significant differences exist. This step was to prevent biases from sample characteristics. This split-sample design and the preliminary equivalence checks are widely recommended in scale development and validation (Worthington & Whittaker, 2006). Sample 1 underwent exploratory analysis in Phase 2, while Sample 2 was for confirmatory analysis in Phase 3. For detailed demographic attributes, psychological characteristics, and t-test outcomes for both datasets, please refer to Tables 1 and 2.

Table 1. *Descriptive information of the two random samples.*

		Sample 1 (n = 277)		Sample 2 (n = 277)	
		Frequency	Percent	Frequency	Percent
Age	18-25	274	98.9	272	98.2
	26-30	1	0.4	3	1.1
	31 and above	2	0.7	2	0.7
Sex	Male	71	25.6	73	26.4
	Female	206	74.4	204	73.6
Grade	Undergraduate students	266	96	265	95.7
	Master's student	8	2.9	6	2.2
	Doctoral student	3	1.1	6	2.2
Major	Humanity	11	4.0	13	4.7

	Sample 1 (*n* = 277)		Sample 2 (*n* = 277)	
	Frequency	Percent	Frequency	Percent
Social science	45	16.2	31	11.2
Natural science	3	1.1	1	0.4
Engineering	5	1.8	6	2.2
Medical	213	76.9	226	81.6

Table 2. *T-test results for demographic and psychological information of the two random samples.*

	Sample	*N*	*M*	*SD*	*t*	*df*	*p*
Age	1	277	/	/	-0.44	552	0.660
	2	277	/	/			
Gender	1	277	/	/	0.143	552	0.886
	2	277	/	/			
Grade	1	277	/	/	-0.559	552	0.576
	2	277	/	/			
Major	1	277	/	/	-1.287	552	0.199
	2	277	/	/			
L2MSS	1	277	3.75	1.02	0.024	552	0.981
	2	277	3.87	1.16			
BFP	1	277	3.08	0.16	1.668	552	0.096
	2	277	3.08	0.16			

Note. During data collection, age, gender, grade, and major were not represented as ordinal or continuous variables; hence, the calculation of mean and standard deviation is not applicable for these characteristics. L2MSS = second language motivational self. BFP = Big Five personality traits.

Procedure and statistical analysis

This study adopted a quantitative correlational research design (Fraenkel *et al.*, 2018), which examines the relationships among variables without researcher manipulation. In the context of scale development, this design enables the systematic analysis of interrelationships among scale items, underlying latent factors, and external constructs, which is essential for establishing the instrument's structural and criterion-related validity.

Before conducting EFA in SPSS 26, we evaluated the suitability of the Sample 1 data. Univariate normality was examined using the Shapiro–Wilk and Anderson–Darling tests. Factorability was assessed via Bartlett's test of sphericity (Bartlett, 1954) and the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy (Kaiser, 1970). Multivariate outliers were screened using Mahalanobis distance, and multivariate normality was evaluated with Mardia's test. Linearity and multicollinearity were examined using the Pearson correlation matrix and bivariate scatterplots among the four L2BS items.

Factor retention in EFA was guided by multiple criteria: the Kaiser criterion (eigenvalues > 1), visual inspection of the scree plot, and Horn's parallel analysis. These methods helped identify the optimal number of factors comprising our scale. Specifically, we considered the number of components with eigenvalues exceeding 1 (Kanyongo, 2006), examined the scree plot for discernible trends (Ledesma *et al.*, 2015), and conducted parallel analysis by comparing eigenvalues from our actual data with those from randomly generated data (Ledesma & Valero-Mora, 2007). During EFA, we utilized oblique (promax) rotation to account for potential correlations among latent constructs (Kieffer, 1998). Factor loadings were categorized as low (< .4), acceptable (.4 to .6), or high (> .6) (Kline, 2014).

Results

Tests of univariate normality (Shapiro–Wilk and Anderson–Darling) indicated significant non-normality for all items ($p < .001$). Mardia's test likewise suggested departures from strict multivariate normality (Mardia's skewness = 122.01, $p < .001$; Mardia's kurtosis = 25.15, $p < .001$). Given EFA's robustness to moderate non-normality with adequate sample sizes ($N = 277$), we proceeded with the analysis while acknowledging this as a methodological limitation.

The findings of Bartlett's test of sphericity and KMO indicated that the dataset was appropriate for further factor analysis. Bartlett's test of sphericity yielded a significant result, with $\chi^2(6) = 616.99$, $p < .001$, indicating that the data were suitable for factor analysis (Field, 2005). The KMO results (KMO = .826) indicated that the data were adequate for subsequent analysis, with KMO exceeding .80 (Field, 2005). A multivariate outlier check using Mahalanobis distance flagged nine cases; because they were not data-entry errors and their removal did not change the factor solution, they were retained in the final analyses. Inter-item Pearson correlations among the four L2BS items ranged from .61 to .73 (Table 3), indicating moderate positive associations without problematic multicollinearity (all $|r| < .80$). Bivariate scatterplots showed approximately elliptical, linear patterns, supporting the assumption of linear relationships among indicators.

Table 3. *Inter-item Pearson correlations among the L2BS items.*

Item	1	2	3	4
L2BS1	1			
L2BS2	0.71	1		
L2BS3	0.73	0.71	1	
L2BS4	0.61	0.68	0.69	1

Note. Entries are Pearson product–moment correlations. All coefficients are significant at $p < .001$. L2BS = Second Language Buoyancy Scale.

The Kaiser criterion (eigenvalues > 1), scree plot, and Horn's parallel analysis converged on a one-factor solution for the L2BS. Only the first factor had an eigenvalue greater than 1, The scree plot showed a clear inflection after the first factor (Figure 2). The parallel analysis indicated that only the first empirical eigenvalue exceeded the corresponding randomly generated eigenvalues (Figure 3). Together, these results supported a one-factor solution for the L2BS. This finding aligns with previous theoretical frameworks (Martin & Marsh, 2008a).

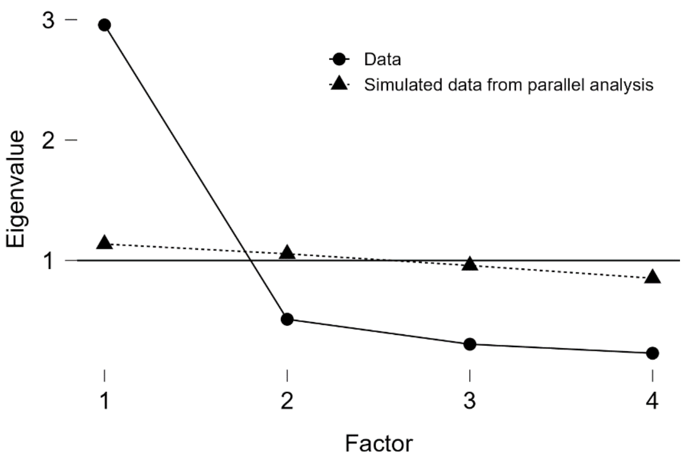


Figure 2. *Scree plot.*

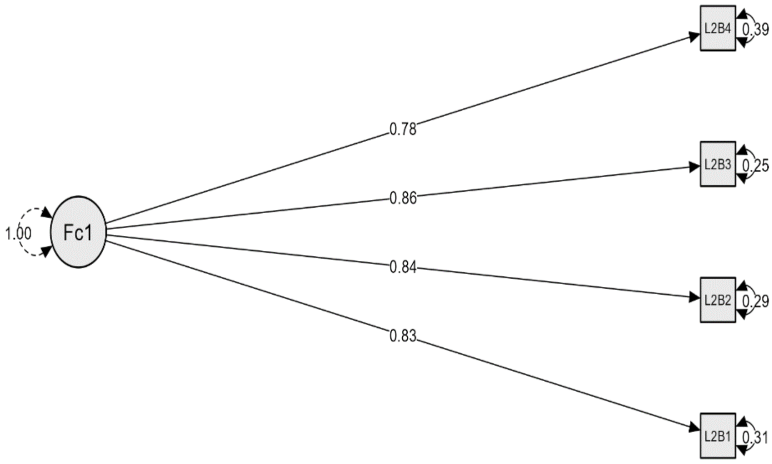


Figure 3. *Model plot.*

EFA with promax rotation confirmed that all four items loaded strongly on a single factor. Standardized factor loadings were all above .65 (Figure 4), and item communalities (h^2) ranged from .61 to .75 (Table 4), indicating that each item shared substantial variance with the common factor and contributed meaningfully to the underlying construct.

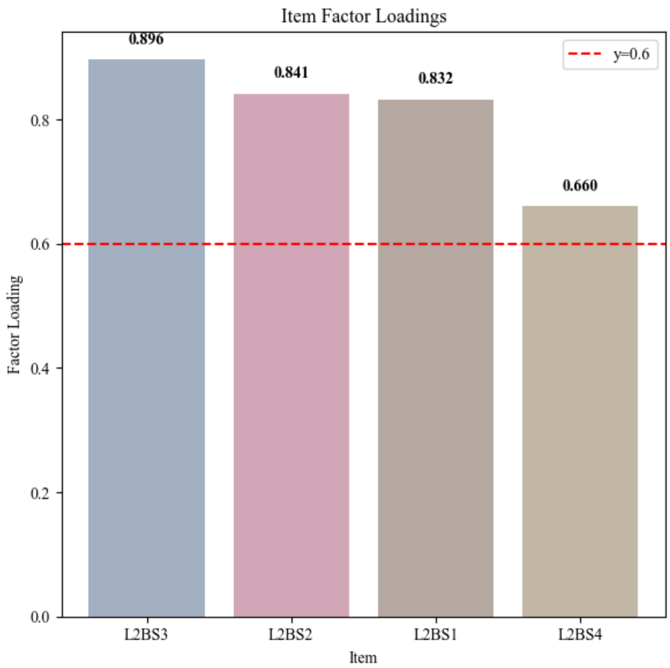


Figure 4. *Item factor loadings.*

Table 4. *Item communalities (h^2) for the L2BS items.*

Item	Communality (h^2)
L2BS1	0.69
L2BS2	0.71
L2BS3	0.75
L2BS4	0.61

Note. Entries are Pearson product–moment correlations. All coefficients are significant at $p < .001$. L2BS = Second Language Buoyancy Scale.

PHASE 3: REFINING L2BS

Phase 3 aimed to confirm the structural validity of the L2BS using CFA. To further assess construct validity, correlational analyses were conducted. In addition, criterion-related validity was examined through hierarchical regression and machine learning methods, using foreign language engagement and enjoyment as outcome variables.

Method

Participants

The dataset used in Phase 3 was the second subset with 277 participants described in the participants subsection under the "INITIAL TEST OF L2BS" heading.

Instruments

The selection of measured constructs in this study was guided by the positive psychology framework in SLA (Dörnyei & Ryan, 2015; MacIntyre *et al.*, 2016), which emphasizes learners' emotional regulation, motivation, and adaptive capacities in language learning contexts.

Academic buoyancy and the newly developed L2BS represent the central constructs of this study. While academic buoyancy reflects students' general ability to manage academic setbacks, L2 buoyancy focuses specifically on language learning contexts, addressing SLA-specific emotional and motivational demands (Yun *et al.*, 2018; Yang *et al.*, 2024).

Growth mindset was selected given its theoretical connection to adaptive motivational patterns and learners' resilience in the face of challenges, which conceptually overlaps with the buoyancy construct. Grit (Duckworth *et al.*, 2007) and conscientiousness (Costa & McCrae, 1992) were included as convergent validity indicators, as both traits have been empirically linked to sustained language learning efforts and long-term achievement (Teimouri *et al.*, 2020).

L2 engagement and L2 enjoyment were included as primary outcome variables because they represent key indicators of learners' sustained involvement and positive affect in the learning process (Dewaele & MacIntyre, 2014; Mercer, 2019). Previous research has demonstrated that buoyant learners tend to report higher levels of enjoyment and engagement, as they possess stronger coping mechanisms to handle academic challenges (Putwain *et al.*, 2015).

These constructs provide a comprehensive framework to evaluate the convergent, discriminant, and predictive validity of the L2BS within a theoretically grounded SLA context.

Academic Buoyancy Scale. The academic buoyancy scale (Martin & Marsh, 2008a) was employed. As the scale was originally in English, two graduate students, both native Chinese speakers, conducted the translation process. Subsequently, the translated items were back-translated into English by the first author, who is also a native Chinese speaker. The entire translation process adhered to the principles outlined for different cultural contexts (Hernández *et al.*, 2020). Participants rated each item on a 7-point Likert scale, ranging from 1 (indicating strong disagreement) to 7 (indicating strong agreement), based on their perceptions. Higher overall scores on the scale refer to better academic buoyancy. The Cronbach's alpha coefficient for the translated version of ABS in this study was 0.91, suggesting high internal consistency. A CFA conducted on the dataset yielded satisfactory fit indices: $\chi^2 = 22.92$, $p < .001$, RMSEA = .14, SRMR = .012, CFI = .99, Tucker–Lewis Index (TLI) = .99, indicating a good model fit. We note as a limitation that we did not implement additional adaptation steps (e.g., expert cultural/structural equivalence review and pilot testing). These are flagged for future work.

While the ABS is not the primary scale in this study and was only adapted linguistically rather than fully culturally validated, it was included as a theoretically relevant criterion measure because it represents the closest established construct within the academic buoyancy framework. However, we acknowledge that its use as a criterion instrument constitutes a methodological limitation, as it may not fully capture domain-specific aspects of L2 learning. In the present study, the ABS was therefore treated as a benchmark rather than a gold-standard criterion measure, and its results were interpreted with caution. Importantly, criterion-related and predictive validity for the L2BS were also evaluated with respect to L2 engagement and enjoyment, rather than relying solely on the ABS as an external criterion.

Language Growth Mindset Scale. To evaluate students' perceptions of their English learning capabilities, we employed the Chinese version of the Growth Mindset Scale, adapted and validated by Yao *et al.* (2021) among Chinese university students. The scale encompasses three items and is structured as a six-point Likert scale, with responses ranging from 1 (strongly disagree) to 6 (strongly agree). In Yao *et al.*'s (2021) validation study, the growth mindset subscale demonstrated good internal consistency (Cronbach's α = .84 - .88 across two time points) and satisfactory construct validity, as confirmed by CFA (CFI > .95, RMSEA ≈ .06, SRMR < .05). In this study, the Cronbach's α for this subscale was .85, indicating acceptable reliability.

Grit Scale. The assessment of individuals' grit in this study utilizes the Chinese adaptation of the Triarchic Model of Grit Scale (TMGS), developed and validated by Datu and Zhang (2021). This scale comprises three core components: passion, perseverance, and adaptability. The original 10-item English version (Datu *et al.*, 2017) was translated into Chinese and back-translated to ensure conceptual equivalence before validation with Chinese vocational college students. Using a 5-point Likert scale, respondents rate the extent to which each statement reflects their personal experiences, with options ranging from "not at all like me" (rated as one) to "very much like me" (rated as five). Prior validation demonstrated adequate reliability (α = .75–.82; ω = .76–.82) and good construct validity (CFA: CFI = .94, TLI = .91, SRMR = .06, RMSEA = .08). In the present study, the overall scale showed acceptable internal consistency (Cronbach's α = .81)

Conscientiousness Scale. To assess conscientiousness, we utilized the conscientiousness subscale of the Chinese version of the NEO Five-Factor Inventory (NEO-FFI), adapted and validated for Chinese adults by Jiang (2020), based on the original inventory by Costa & McCrae (1992). This subscale comprises 12 items rated on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). The conscientiousness subscale of the Chinese NEO-FFI has been widely used in psychological research in China, and its reliability and validity have been repeatedly confirmed (Jiang, 2020); therefore, further convergent validity analyses were not conducted, given the extensive prior validation of the instrument. When factor loadings exceed .50, composite reliability values are above .60, and AVE values exceed .50, the convergent validity of the scale can be considered acceptable. Prior validation with Chinese adult participants reported good reliability (Cronbach's α ≈ .80) and satisfactory construct validity of the conscientiousness dimension. In this study, the subscale also demonstrated acceptable internal consistency (Cronbach's α = .80).

L2 Engagement Scale. L2 engagement was assessed using the Foreign Language Engagement scale developed and validated by Li *et al.* (2023) in the Chinese context. The scale contains 15 items across four sub-dimensions: cognitive, emotional, behavioral, and social engagement. Items are rated on a 5-point Likert scale (1 = completely disagree to 5 = agree). Prior validation reported good reliability (Cronbach's α = .90 for the total scale; subscales ranging from .46 to .89) and satisfactory construct validity supported by CFA (e.g., CFI ≈ .95, RMSEA ≈ .06). In the present study, the scale demonstrated acceptable internal consistency (Cronbach's α = .84).

L2 Enjoyment Scale. The 11-item Chinese Version of the Foreign Language Enjoyment Scale (CFLES) developed by Li *et al.* (2018) was employed to gauge the extent of enjoyment derived from L2 learning. The original Foreign Language Enjoyment Scale (Dewaele & MacIntyre, 2014, 2016) was first translated into Chinese following a rigorous translation and back-translation procedure by bilingual experts. Subsequent revisions were made based on feedback from teachers and students before large-scale validation. The CFLES was validated with two independent samples of over 2,000 Chinese high school students, and a three-factor structure (FLE-Private, FLE-Teacher, FLE-Atmosphere) was confirmed. Employing a conventional 5-point Likert scale, participants indicated their level of agreement across a range from "1" indicating strongly disagree to "5" indicating strongly agree. Psychometric analyses demonstrated good reliability and validity: Cronbach's α = .83 (global scale), with subscale alphas of .79 (Private), .90 (Teacher), and .78 (Atmosphere), and CFA results showing excellent fit indices (CFI = .975, TLI = .967, RMSEA = .041). In this study, the Cronbach's α was .82.

Statistical analysis

Structural Validity. CFA was conducted using maximum likelihood estimation in AMOS 23. In addition to data-level assumptions, we examined structural assumptions by inspecting modification indices and standardized residuals. No substantial or theoretically justified correlated error terms were detected. Therefore, no error covariances were added. These results support local independence and error, factor independence in the one-factor CFA model.

Reliability, including internal consistency, was measured with Cronbach's alpha and McDonald's omega.

Construct Validity. Construct validity was assessed through convergent validity (using AVE) and discriminant validity (correlating L2 buoyancy scores with academic buoyancy, language mindsets, grit, and conscientiousness).

Criterion-related Validity. We evaluated the predictive capabilities of L2BS and ABS for L2 engagement and enjoyment using hierarchical regression and machine learning methods in Python 3.12, highlighting L2BS's suitability for L2 learning contexts.

Hierarchical regression was used because it allows for the sequential evaluation of predictor variables, enabling the examination of the incremental contribution of L2 buoyancy dimensions beyond demographic factors and general academic buoyancy. This method is particularly suitable for testing incremental validity, as it provides clear evidence of whether L2BS explains additional variance in learning outcomes after accounting for known predictors (Cohen *et al.*, 2003; Aiken & West, 1991). By entering variables in successive blocks, hierarchical regression facilitates a structured evaluation of the extent to which unique variance in L2 engagement and enjoyment is attributable to L2-specific buoyancy components.

This study also employed three models (random forests, XGBoost, and LightGBM) in ML to compare L2BS and ABS and examine L2BS's adaptability in L2 environments from a localized perspective.

A Random Forest (RF) model is an ensemble learning method that combines multiple decision trees to produce a strong, reliable predictive model (Breiman, 2001). It is valued for its simplicity, efficiency, and ability to manage large, high-dimensional datasets. The algorithm constructs numerous decision trees using bootstrap samples from the training data and then averages their outputs to generate predictions. This ensemble approach enhances generalization and reduces the risk of overfitting. The predictive performance of the RF model can be mathematically represented in Equation 1:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \tag{1}$$

Note. T_i: prediction of the i^{th} tree; N: number of trees.

Gradient Boosted Decision Trees (GBDT) is an ensemble learning method that builds trees in a sequential manner, where each new tree is trained to correct the errors made by the previous ones. Two prominent implementations of this approach are XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke *et al.*, 2017), both recognized for their high efficiency and scalability.

XGBoost enhances optimization by using Newton's method, which relies on a second-order Taylor expansion of the loss function to improve convergence accuracy. Moreover, it incorporates a regularization term in the objective function (Equation 2) to manage model complexity and reduce the risk of overfitting:

$$\mathrm{Obj}(t) \approx \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y_1^{(t-1)}}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{2}$$

Note. L: the loss function; g_i: the first derivative of the loss function; h_i: the second derivative of the loss function; f_t: the output of the t^{th} tree; $\Omega(f_t)$: the regularization term.

LightGBM prioritizes reducing training time while maintaining accuracy. It introduces Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to manage large datasets efficiently. GOSS retains instances with large gradients, and EFB groups mutually exclusive features, thereby reducing computational complexity. In contrast to XGBoost, LightGBM grows trees in a leaf-wise manner, selecting splits that result in the greatest reduction in loss (Equation 3):

$$\mathrm{Gain} = \mathrm{Loss}_{\mathrm{before}} - \mathrm{Loss}_{\mathrm{after}} \tag{3}$$

Overall, RF provides a simple yet robust ensemble learning method, while XGBoost and LightGBM enhance GBDT with advanced optimization strategies and improved efficiency. These models are well-suited for complex predictive tasks.

To evaluate model performance, four metrics were used: accuracy, precision, recall, and F1 score. Together, they offer a well-rounded assessment by capturing overall prediction correctness (accuracy) and the trade-off between precision and recall.

Accuracy, defined as the ratio of correct predictions (both true positives and true negatives) to total cases, is calculated as in Equation 4:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+FN+FP+TN}} \tag{4}$$

Note. TP (True Positives): correctly predicted positive cases; TN (True Negatives): correctly predicted negative cases; FP (False Positives): negative cases incorrectly labeled as positive; FN (False Negatives): positive cases incorrectly labeled as negative.

The F1-score represents the harmonic mean of precision and recall, emphasizing their balance (Equation 5):

$$\text{F1-Score} = 2 \cdot \frac{\text{precision-recall}}{\text{precision+recall}} \tag{5}$$

Equations 6 and 7 define the precision and recall:

$$\mathrm{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$\mathrm{Recall} = \frac{TP}{TP+FN} \tag{7}$$

Precision measures how many of the instances the model labeled as positive are actually positive. Recall measures how many of the truly positive instances the model successfully identified.

RESULTS

Structural Validity

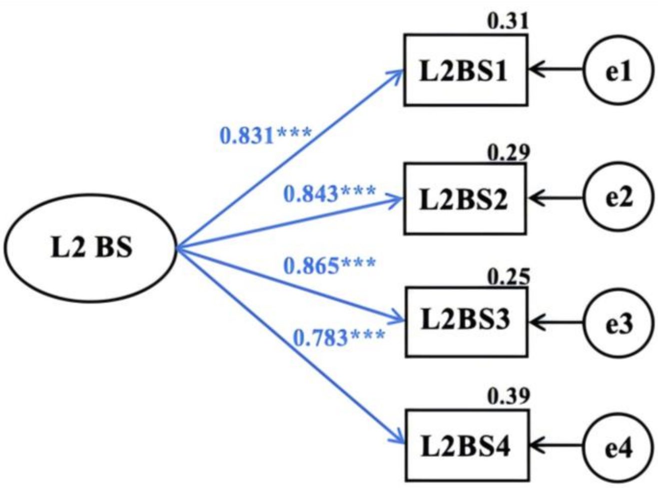
In this step, a CFA was conducted to assess the proposed model of the 4-item L2BS. Prior to conducting CFA, we tested the statistical assumptions. Mardia's multivariate skewness (2.31, p = .119) and kurtosis (18.75, p = .087) were non-significant, indicating approximate multivariate normality. Mahalanobis distance values were compared to the χ^2 distribution with df = 4 at p < .001 (critical value = 18.47); no cases exceeded this threshold, suggesting the absence of multivariate outliers. Missing data were minimal (<1%) and handled using full information maximum likelihood (FIML). All items were measured on a 7-point Likert scale and treated as continuous indicators, a common approach when categories exceed five (Rhemtulla *et al.*, 2012). Robust maximum likelihood (MLR) estimation with Satorra–Bentler correction was used to account for non-normality. The sample size (N = 277) exceeded the recommended minimum of 200 for CFA (Kline, 2005), ensuring stable parameter estimation.

The fit indices indicated a satisfactory model fit, with a Root Mean Square Error of Approximation (RMSEA) of .096, which falls below the recommended threshold of 0.1. Additionally, the CFI of .992 and the TLI of .977 both suggested a close fit, surpassing the conventional cutoff of .95 (Kline, 2005). Importantly, the Standardized Root Mean Square Residual (SRMR) was .041, well below the recommended threshold of .08, further supporting the adequacy of the model.

The Satorra–Bentler chi-square test returned a significant result ($\chi^2(2) = 7.116$; p < .05), yet the chi-square to degrees of freedom ratio (χ^2/df = 3.58) surpassed the recommended threshold of 2 (Byrne, 1989). However, it's important to acknowledge that both the chi-square value and the χ^2/df ratio can be influenced by sample size and are susceptible to Type I errors (Kenny, 2020). Given the relatively large sample size utilized in this analysis, this outcome is deemed acceptable.

All factor loadings were high and statistically significant, indicating that each of the four items loaded strongly onto the latent construct. A visual representation of the CFA model with standardized factor loadings is presented in Figure 5 to illustrate the measurement structure of the 4-item L2BS.

Inspection of modification indices did not reveal any substantial correlated residuals or cross-loadings that would be theoretically warranted, and standardized residuals were small, providing additional support for the adequacy of the specified one-factor model and the assumption of independent error terms.



Note. *** p < .001.

Figure 5. *CFA result of L2BS.*

Reliability

In terms of internal consistency, the scale exhibited a Cronbach's alpha of α = .898 and McDonald's omega of ω = .898, indicating a strong level of internal consistency (Cronbach, 1951; McDonald, 1999).

Construct Validity

Convergent validity was assessed by calculating the AVE of the single factor, yielding an AVE value of .689, which exceeded the recommended minimum threshold of .50 suggested by Franke and Sarstedt (2019). Discriminant validity was established by examining the correlations between L2 buoyancy and other psychological traits. The small-to-moderate correlations observed between L2 buoyancy, academic buoyance, grit, growth mindset, and conscientiousness confirmed that the construct measured by L2 buoyancy is distinct from other psychological variables (Figure 6).

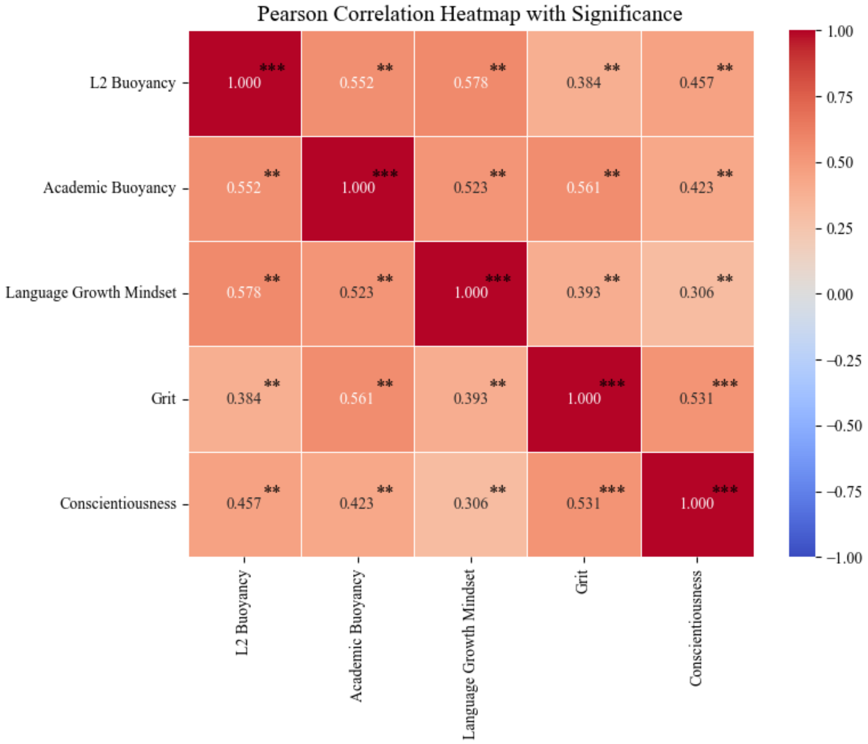


Figure 6. *The correlation (Pearson's r) results between L2 buoyancy, academic buoyance, grit, growth mindset, and conscientiousness.*

Criterion-Related Validity

Hierarchical regression analysis

To examine the predictive validity of the newly developed L2BS, hierarchical regression analyses were conducted for two dependent variables: L2 engagement and L2 enjoyment, and the results are presented in Table 5. Prior to analyses, all standard assumptions for regression were tested and satisfied: linearity was supported by scatterplots; residual normality showed Shapiro–Wilk departures (engagement: $W = 0.911$, $p < .001$; enjoyment: $W = 0.951$, $p < .001$) but Q–Q plots indicated approximate normality and, given $N = 277$, any bias is negligible (Schmidt & Finan, 2018). Tests of homoscedasticity supported the assumption, with non-significant Breusch–Pagan statistics ($\chi^2 = 0.09$, $p = .958$ for L2 engagement; $\chi^2 = 3.71$, $p = .156$ for L2 enjoyment). Errors were independent (Durbin–Watson = 1.89 - 2.11; Field, 2024). Multicollinearity was low (both VIFs = 1.44; tolerance > .70), and no influential observations were detected (maximum Cook's D = 0.28 and 0.35, both < 1). These diagnostics support the validity of the reported hierarchical regressions.

Table 5. *Hierarchical regression results.*

Dependent Variable	Model		R^2	F	p	B	t	p
L2 Engagement	1	(Intercept)	.199	16.894	<.001	2.400	10.232	< .001
		Age				-0.187	-1.101	.272
		Grade				0.116	0.977	.329
		Major				0.019	0.826	.409
		ABuo				0.168	8.067	< .001
	2	(Intercept)	.372	32.082	<.001	2.022	9.505	< .001
		Age				-0.027	-0.175	.861
		Grade				0.020	0.192	.848
		Major				-0.003	-0.169	.866
		ABuo				0.061	2.725	.007
		L2Buo				0.202	8.635	< .001
L2 Enjoyment	1	(Intercept)	.381	41.809	<.001	1.842	5.987	< .001
		Age				-0.314	-1.408	.160
		Grade				0.160	1.025	.306
		Major				0.015	0.509	.611
		ABuo				0.349	12.781	< .001
	2	(Intercept)	.541	63.946	<.001	1.298	4.786	< .001
		Age				-0.083	-0.429	.668
		Grade				0.022	0.164	.870
		Major				-0.017	-0.647	.518
		ABuo				0.195	6.863	< .001
		L2Buo				0.291	9.737	< .001

Note. ABuo stands for Academic Buoyancy. L2Buo stands for L2 Buoyancy.

For L2 engagement, in Model 1, demographic variables and academic buoyancy (ABuo) explained 19.9% of variance ($R^2 = .199$, $F(4, 295) = 16.894$, $p < .001$). Adding L2 Buoyancy (L2Buo) in Model 2 significantly increased explained variance by 17.3% ($\Delta R^2 = .173$), resulting in a total R^2 of 0.372 ($F(5, 294) = 32.082$, $p < .001$). Among predictors, both ABuo ($B = 0.061$, $t = 2.725$, $p = .007$) and L2Buo ($B = 0.202$, $t = 8.635$, $p < .001$) significantly predicted L2 engagement.

For L2 enjoyment, Model 1 explained 38.1% of variance ($R^2 = .381$, $F(4, 295) = 41.809$, $p < .001$), with ABuo showing a significant effect ($B = 0.349$, $t = 12.781$, $p < .001$). The inclusion of L2Buo in Model 2 further increased explained variance by 16.0% ($\Delta R^2 = .160$), yielding a total R^2 of .541 ($F(5, 294) = 63.946$, $p < .001$). In this model, both ABuo ($B = 0.195$, $t = 6.863$, $p < .001$) and L2Buo ($B = 0.291$, $t = 9.737$, $p < .001$) remained significant predictors.

These findings indicate that L2 buoyancy explains substantial additional variance in both L2 engagement and enjoyment beyond that accounted for by general academic buoyancy alone, supporting the incremental validity of the L2BS.

Machine learning predictive validity

A machine learning classification task was conducted to predict two target outcomes: L2 enjoyment and L2 engagement. During preprocessing, both continuous scores were discretized into three ordered categories (low, medium, high) using tertile cut-points to enable multiclass classification. This discretization approach is commonly used in educational prediction research to enhance interpretability while retaining essential information for modeling (Dewaele & Alfawzan, 2018; Li *et al.*, 2020). Predictor variables included the L2BS and sub-dimensions of ABS.

The dataset was split into 80% for training and 20% for testing for model evaluation. Models were implemented in Python using Random Forest, XGBoost, and LightGBM. Appendix 4 outlined the parameter settings for each model.

For L2 Enjoyment prediction (Table 6), all L2BS-based models outperformed ABS-based models. The highest performance was achieved by the XGBoost model with L2BS, yielding an accuracy of 64.29%, precision of 71.60%, recall of 64.29%, and F1-score of 66.72%. In contrast, the best model using ABS (Random Forest) reached lower performance: accuracy of 60.71%, precision of 61.87%, recall of 60.71%, and F1-score of 60.97%.

Table 6. *Model performance comparison in predicting L2 enjoyment using L2BS and ABS.*

Predicting L2 Enjoyment: Model Performance Comparison (Weighted Average)								
	Performance of L2BS				Performance of ABS			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Random Forest	62.50%	71.07%	62.50%	65.33%	60.71%	61.87%	60.71%	60.97%
XGBoost	64.29%	71.60%	64.29%	66.72%	55.36%	61.03%	55.36%	57.42%
LightGBM	62.50%	68.68%	62.50%	64.09%	57.14%	61.07%	57.14%	58.38%

For L2 Engagement prediction (Table 7), models based on L2BS also demonstrated superior predictive accuracy. Both Random Forest and LightGBM models with L2BS achieved identical and the highest scores: accuracy of 73.21%, precision of 74.64%, recall of 73.21%, and F1-score of 73.04%. In comparison, the best ABS-based model (LightGBM) achieved 71.43% accuracy and a lower F1-score of 70.08%

Table 7. *Model performance comparison in predicting L2 engagement using L2BS and ABS.*

Predicting L2 Engagement: Model Performance Comparison (Weighted Average)								
	Performance of L2BS				Performance of ABS			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Random Forest	73.21%	74.64%	73.21%	73.04%	64.29%	65.31%	64.29%	63.99%
XGBoost	73.21%	73.89%	73.21%	73.05%	69.64%	67.60%	69.64%	68.30%
LightGBM	73.21%	74.64%	73.21%	73.04%	71.43%	69.39%	71.43%	70.08%

Therefore, these results from both regression and machine learning analyses provide converging evidence that the L2BS offers stronger predictive validity than the traditional ABS in forecasting key L2 learning outcomes, particularly L2 engagement and enjoyment. The final version of L2BS was presented in Appendix 5.

DISCUSSION

Summary of Findings

The primary goal of this research was to provide preliminary evidence of the psychometric validity of the L2BS using data from two distinct groups of Chinese college students. Overall, the findings indicated that the 4-item L2BS is psychometrically robust and offers a valid measure. Therefore, the L2BS was found to be a reliable tool for assessing L2 buoyancy, facilitating the exploration of various learning styles in SLA.

Psychometric Properties of L2BS

The first research question of the study focused on elucidating the factor structure of the L2BS, a fundamental aspect in understanding its composition. During this exploration, the L2BS was found to be unidimensional. Although we did not test more complex multidimensional structures in this preliminary validation, we adopted a single-factor model because L2 buoyancy has been conceptualized as a unitary capacity to cope with routine language-learning challenges. Future research could examine the viability of bifactor or multidimensional models to further probe whether more fine-grained facets of L2 buoyancy can be distinguished. This single-factor solution suggests that the four L2BS items tap a common underlying capacity to deal with everyday setbacks in L2 learning rather than reflecting several qualitatively different dimensions (e.g., affective, behavioral, or cognitive buoyancy). In line with Martin and Marsh's (2008a) conceptualization of academic buoyancy, our findings indicate that a single latent construct can meaningfully summarize language learners' responses to routine L2 challenges.

The second research question sought to determine the validity and reliability of L2BS as an assessment tool. The measurement model proposed for L2BS showed a satisfactory fit. The good model fit indices (e.g., CFI/TLI above conventional cut-off values and RMSEA/SRMR within acceptable ranges) indicate that the hypothesized L2 buoyancy factor reproduces the observed covariance structure well. This provides empirical support that the theoretical structure of L2 buoyancy - as a unitary, situation-specific protective factor - is consistent with how students actually respond to the items.

Furthermore, L2BS exhibited robust reliability and validity, with acceptable internal consistency. High internal consistency further implies that the items coherently capture students' tendency to "bounce back" from typical L2 difficulties, which is central to the definition of L2 buoyancy. The internal consistency and composite reliability coefficients observed in this study are comparable to or slightly higher than those reported for the original ABS and other motivation-related scales in SLA (e.g., grit, growth mindset), despite the L2BS's shorter length. This pattern is noteworthy because reliability typically decreases as the number of items decreases. The relatively strong coefficients in our Chinese college sample suggest that the four items are highly homogeneous and that L2 buoyancy is a well-defined construct in this context. At the same time, the high inter-item homogeneity may also indicate limited content diversity, which reinforces the need to broaden the item pool in future research. Moreover, given that our sample is relatively academically homogeneous, it will be important to examine whether similar reliability levels can be replicated in more diverse learner populations.

Evidence supporting the construct validity of the scale was also established. Given the scale's concise nature and limited number of items, it can be readily incorporated into research studies examining multiple variables. It is anticipated that the introduction of this measure will enhance comprehension of the nomological network within which L2 buoyancy is situated. At the same time, the brevity of the four-item scale inevitably constrains its structural representational power. While the items capture core aspects of bouncing back from everyday L2 difficulties, they may not fully cover the broader range of cognitive, emotional, and behavioral strategies that learners employ in response to language-learning setbacks. Thus, the present L2BS should be regarded as a parsimonious index of L2 buoyancy rather than an exhaustive representation of all possible facets of the construct. Future studies could expand the item pool to sample a wider range of situations and reactions, and then use more sophisticated models (e.g., bifactor CFA or ESEM) to examine whether multiple facets of L2 buoyancy can be distinguished while retaining a general factor.

Comparison with Existing Instruments and Factor Structure

The third research question explored whether L2BS could predict L2 learning-related outcomes using both regression and machine learning methods to assess its relevance in the L2 context. Regression results showed that students with higher L2 buoyancy were more engaged in class activities and more likely to enjoy language learning than those with lower buoyancy, aligning with previous findings (Yun *et al*., 2018). These results support the reliability of L2BS as an indicator of behavioral patterns among L2 learners. Both regression and machine learning analyses further demonstrated that L2BS outperformed ABS in predicting L2 enjoyment and engagement, highlighting its value as a predictive tool in language learning environments. These findings suggest that a domain-specific buoyancy measure may capture language learners' everyday coping resources more precisely than a domain-general academic buoyancy scale. From a control-value perspective, students who are better able to "bounce back" from routine L2 difficulties may experience higher perceived control and more positive achievement emotions (e.g., enjoyment), which in turn support sustained behavioral engagement in language classes.

In addition to predictive validity, further validity evidence is needed to establish the cross-cultural applicability of the L2BS. Specifically, future studies should conduct measurement invariance tests (MG-CFA) across relevant groups to ensure configural, metric, scalar, and strict invariance. Moreover, differential item functioning (DIF) analyses are recommended to detect potential item bias across sociocultural and environmental contexts. Such analyses will complement the present machine learning results by providing more substantial evidence that the L2BS functions equivalently and fairly across diverse learner populations.

Furthermore, our analysis revealed a positive correlation between L2 buoyancy and all other psychological traits (i.e., language growth mindsets, grit, and conscientiousness) in language learners. In particular, L2 buoyancy demonstrated a stronger association with the language-specific trait of language growth mindset than with the other two broader personality traits (i.e., grit and conscientiousness). This pattern emphasizes the distinct influence of L2 buoyancy within the realm of language acquisition, supporting its unique role in language learning. The stronger association between L2 buoyancy and language growth mindset than with grit or conscientiousness indicates that how students interpret and appraise L2 setbacks (e.g., as opportunities to improve versus signals of low ability) may be more central to buoyancy than their broader personality traits. This pattern aligns with motivational theories in SLA, which emphasize the role of domain-specific beliefs and appraisals in shaping learners' persistence and emotional responses.

Beyond these correlational patterns, the present findings also point to several plausible mediating processes that future research could test more explicitly. For example, L2 buoyancy may foster more adaptive self-regulation (e.g., strategic responses to errors), more positive motivational beliefs about the malleability of L2 ability, and higher levels of cognitive-emotional resilience in the face of classroom setbacks. Longitudinal and mediation models could examine whether such processes statistically mediate the links between L2 buoyancy and downstream outcomes such as L2 engagement, enjoyment, and achievement.

Contribution of Machine Learning Analyses

Beyond traditional regression analyses, the ML models (e.g., Random Forest, XGBoost, and LightGBM) provided additional evidence for the predictive utility of the L2BS. Across algorithms, L2 buoyancy consistently emerged as one of the most important predictors of L2 enjoyment and engagement, and models including the L2BS achieved higher predictive accuracy than those relying solely on the ABS or other psychological traits. These findings indicate that the psychometrically sound L2BS does not merely reflect a statistically coherent construct (as shown by the CFA) but also captures variance that is highly informative for predicting meaningful language-learning outcomes.

In the present study, the ML analyses primarily used the total L2BS score as a predictor, and thus speak most directly to the predictive validity of the scale. As such, evidence for the unidimensional structure of the L2BS still rests mainly on the EFA and CFA results, whereas the ML findings complement these analyses by demonstrating that the overall L2 buoyancy score is highly informative for predicting L2 engagement and enjoyment.

CONCLUSION and IMPLICATIONS

In this study, we developed and validated a measure of L2 buoyancy, the L2BS, based on the 4-item ABS, demonstrating it to be a valid and reliable instrument. This measure is recommended for research examining L2 buoyancy among adolescents or adults in an L2 classroom context.

The present findings have several implications for language teaching practice. Because the L2BS is brief and psychometrically sound, teachers and practitioners can use it as a screening tool to identify students who are more vulnerable to everyday L2 setbacks (e.g., test disappointments, classroom mistakes). Repeated administrations across a semester may help monitor changes in students' buoyancy and evaluate the impact of pedagogical interventions.

Given that L2 buoyancy was strongly related to L2 enjoyment and engagement, interventions that explicitly aim to strengthen students' capacity to "bounce back" from routine difficulties may indirectly enhance their participation and positive emotions in class. For example, teachers can provide structured opportunities to reflect on setbacks, model adaptive coping strategies, and incorporate growth-oriented feedback that emphasizes effort, strategy use, and learning from mistakes. Such practices may foster a classroom climate in which temporary failures are normalized and viewed as stepping stones in L2 development.

Furthermore, the stronger association between L2 buoyancy and language-specific growth mindset suggests that classroom activities that target students' beliefs about the malleability of L2 ability (e.g., discussing scientific evidence on language learning, sharing narratives of successful learners who initially struggled) may be particularly effective in boosting buoyancy. The L2BS can serve as a practical tool for teachers to identify students who might benefit most from such mindset- and resilience-building interventions.

Limitations and Future Directions

This study is not without its limitations. First, the L2BS currently consists of only four items. Although this ultra-brief format is practical for screening and monitoring, it may not fully capture the complexity and finer facets of L2 buoyancy, especially at extreme trait levels. The scale should therefore be viewed as a parsimonious index rather than a comprehensive diagnostic tool. Future research should expand the item pool and compare the short and long forms on structural validity, measurement invariance, and predictive accuracy, ideally using advanced models such as bifactor CFA or ESEM to test whether distinct facets of L2 buoyancy can be identified while retaining a general factor. Second, the data were collected from Chinese college students, a sample skewed toward medical majors, limiting the generalizability of the findings to other disciplines, educational levels, and sociocultural contexts. In addition, we did not examine measurement invariance or DIF across subgroups (e.g., gender, major, proficiency level) because subgroup sizes were highly unbalanced, which would likely have produced unstable estimates and inflated Type I and Type II errors. Moreover, the ABS was translated and back-translated; we did not implement all recommended adaptation procedures (e.g., expert review for cultural/structural equivalence, small-scale piloting, formal bilingual equivalence testing, or statistical linking), which may limit cross-language comparability. Future studies should use larger and more balanced samples to conduct MG-CFA and DIF analyses, apply systematic adaptation procedures, and include cross-cultural or bilingual designs to strengthen cross-group and cross-language validity evidence. Finally, the study relied on L2 enjoyment and engagement as outcome variables and did not include objective indicators of language achievement (e.g., course grades or standardized test scores), limiting the evidence for predictive validity to self-reported psychological outcomes. Subsequent research should incorporate objective performance measures to provide more direct evidence of the L2BS's capacity to predict L2 achievement. Finally, the ML models were trained and tested on a single dataset, raising concerns about potential overfitting and limited generalizability. In the machine learning analyses, continuous L2 engagement and enjoyment scores were discretized into three categories (low, medium, high) to enable multiclass classification. Although this approach improves interpretability, it inevitably entails some information loss and may attenuate predictive precision. Future work should complement such classification models with algorithms that directly model continuous outcomes (e.g., regression-based ML models) and evaluate whether the L2BS shows comparable or stronger predictive performance under these settings. Future ML work should therefore implement more rigorous validation strategies, such as k-fold cross-validation and replication on independent samples, to reduce the risk of overfitting and to improve the generalizability and transparency of the predictive models.

Acknowledgments

This study was supported by the [University of Macau] under Grant [MYRG-GRG2024-00137-FED].

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: University of Macau, SSHRE24-APP024-FED.

Data Availability Statement

Research data is available and can be sent via email upon request

Informed Consent

Written informed consent was obtained from the participants who participated in this study.

Contribution of Authors

Kenan Gao: Conceptualization, Investigation, Resources, Visualization, Software, Methodology, Data curation, Writing-original draft. **Juan Zhang**: Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Yihui Wang** & **Wei He**: Formal analysis, Writing-original draft. **Jianhong Mo**: Writing – review & editing.

Orcid

Kenan Gao https://orcid.org/0009-0004-7476-9514
Juan Zhang https://orcid.org/0000-0002-7052-1093
Yihui Wang https://orcid.org/0000-0002-3221-3313
Wei He https://orcid.org/0000-0001-7786-2715
Jianhong Mo https://orcid.org/0000-0003-1481-3099

REFERENCES

Aiken, L.S., West, S.G., & Reno, R.R. (1991). *Multiple regression: Testing and Interpreting Interactions*. Sage.

Bartlett, M.S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *16*(2) 296-298.

Bayrami, M., Heshmati, R., Mohammadpour, V., Gholamzadeh, M., Hasanloo, H.O., & Moslemifar, M. (2012). Happiness and willingness to communicate in three attachment styles: a study on college students. *Procedia – Social and Behavioral Sciences*, *46*, 294-298. https://doi.org/10.1016/j.sbspro.2012.05.109

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Clark, L.A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development. In A.E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4th ed., pp. 187–203). American Psychological Association. https://doi.org/10.1037/14805-012

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Collie, R.J., Martin, A.J., Malmberg, L.E., Hall, J., & Ginns, P. (2015). Academic buoyancy, student's achievement, and the linking role of control: A cross-lagged analysis of high school students. *British journal of educational psychology*, *85*(1), 113-130. https://doi.org/10.1111/bjep.12066

Collie, R.J., Perry, N.E., & Martin, A.J. (2017). School context and educational system factors impacting educator stress. In T.M. McIntyre, S.E. McIntyre, & D.J. Francis (Eds.), *Educator stress: An occupational health perspective* (pp. 3–22). Springer International Publishing/Springer Nature. https://doi.org/10.1007/978-3-319-53053-6_1

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Costa, P.T., & McCrae, R.R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*(1), 5-13. https://doi.org/10.1037/1040-3590.4.1.5

Datu, J.A.D., & Yang, W. (2018). Psychometric validity and gender invariance of the academic buoyancy scale in the Philippines: A construct validation approach. *Journal of Psychoeducational Assessment*, *36*(3), 278-283. https://doi.org/10.1177/0734282916674423

Datu, J.A.D., Yuen, M., & Chen, G. (2017). Development and validation of the Triarchic Model of Grit Scale (TMGS): Evidence from Filipino undergraduate students. *Personality and Individual Differences*, *114*, 198-205. https://doi.org/10.1016/j.paid.2017.04.012

Datu, J.A.D., & Zhang, J. (2021). Validating the Chinese version of triarchic model of grit scale in technical–vocational college students. *Journal of Psychoeducational Assessment*, *39*(3), 381-387. https://doi.org/10.1177/0734282920974813

Dewaele, J.M., & Alfawzan, M. (2018). Does the effect of enjoyment outweigh that of anxiety in foreign language performance?. *Studies in second language learning and teaching*, *8*(1), 21-45.

Dewaele, J.M., & MacIntyre, P.D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, *4*(2), 237-274.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. Routledge.

Dweck, C.S. (2013). *Self-theories: Their role in motivation, personality, and development*. Psychology press.

Farrington, C.A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T.S., Johnson, D.W., & Beechum, N.O. (2012). Teaching adolescents to become learners. *The role of noncognitive factors in shaping school performance: A critical literature review*. University of Chicago Consortium on Chicago School Research.

Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. Sage publications limited.

Field, A.P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary?. *Psychological Methods*, *10*(4), 444-467. https://doi.org/10.1037/1082-989X.10.4.444

Franke, G., & Sarstedt, M. (2019). Heuristics versus statistics in discriminant validity testing: a comparison of four procedures. *Internet Research*, *29*(3), 430-447. https://doi.org/10.1108/IntR-12-2017-0515

Fredrickson, B.L. (2001). The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. *American Psychology*, *56* (3), 218-226. https://doi.org/10.1037/0003-066X.56.3.218

Fu, L. (2024). Social support in class and learning burnout among Chinese EFL learners in higher education: Are academic buoyancy and class level important? *Current Psychology*, *43*(7), 5789-5803. https://doi.org/10.1007/s12144-023-04778-9

Gao, K., Zhang, J., Wang, Y., Wu, C., & He, W. (2025). Development and validation of the L2 grit scale in collective cultural context (L2GSC): Evidence based on machine learning methods. *Journal of China Computer-Assisted Language Learning*. https://doi.org/10.1515/jccall-2024-0018

Granziera, H., Liem, G.A.D., Chong, W.H., Martin, A.J., Collie, R.J., Bishop, M., & Tynan, L. (2022). The role of teachers' instrumental and emotional support in students' academic buoyancy, engagement, and academic skills: A study of high school and elementary school students in different national contexts. *Learning and Instruction*, *80*, Article 101619. https://doi.org/10.1016/j.learninstruc.2022.101619

Gu, J., Zhan, P., Liu, J., & Wang, J.-L. (2023). Strength-based parenting and academic buoyancy: a short-term longitudinal chain mediation model. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*, *43*(8), 6753-6760. https://doi.org/10.1007/s12144-023-04892-8

Hejazi, E., & Abbasi, F. (2021). The Effect of Perceived Parental Relationships, Teacher-Student Relationship and Personal Best Goals on Academic Buoyancy. *Quarterly of Applied Psychology*, *15* (2), 179-205. https://doi.org/10.52547/apsy.2021.216011.0

Hernández, A., Hidalgo, M.D., Hambleton, R.K., & Gómez Benito, J. (2020). International test commission guidelines for test adaptation: A criterion checklist. *Psicothema*, *32*(3), 390-398. https://doi.org/10.7334/psicothema2019.306

Hirvonen, R., Putwain, D.W., Määttä, S., Ahonen, T., & Kiuru, N. (2020). The role of academic buoyancy and emotions in students' learning-related expectations and behaviours in primary school. *British Journal of Educational Psychology*, *90*(4), 948-963. https://doi.org/10.1111/bjep.12336

Hirvonen, R., Yli-Kivistö, L., Putwain, D.W., Ahonen, T., & Kiuru, N. (2019). School-related stress among sixth-grade students–Associations with academic buoyancy and temperament. *Learning and Individual Differences*, *70*, 100-108. https://doi.org/10.1016/j.lindif.2019.01.012

Hoferichter, F., Hirvonen, R., & Kiuru, N. (2021). The development of school well-being in secondary school: High academic buoyancy and supportive class-and school climate as buffers. *Learning and Instruction*, *71*, Article 101377. https://doi.org/10.1016/j.learninstruc.2020.101377

Jahedizadeh, S., Ghonsooly, B., & Ghanizadeh, A. (2019). Academic buoyancy in higher education: Developing sustainability in language learning through encouraging buoyant EFL students. *Journal of Applied Research in Higher Education*, *11*(2), 162-177. https://doi.org/10.1108/JARHE-04-2018-0067

Jiang, Y. (2020). *A study on the multi-level impact of personality traits and organizational innovation situations on employee career growth* [Unpublished doctoral dissertation]. University of Science and Technology of China.

Kaiser, H.F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401-415. https://doi.org/10.1007/BF02291817

Kanyongo, G.Y. (2006). The influence of reliability on four rules for determining the number of components to retain. *Journal of Modern Applied Statistical Methods*, *5*, 332-343. https://doi.org/10.22237/jmasm/1162353960

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3149-3157). Curran Associates Inc. https://dl.acm.org/doi/10.5555/3294996.3295074

Kenny, D.A. (2024, October 6). *Measuring model fit*. http://davidakenny.net/cm/fit.htm

Kieffer, K.M. (1998, November 4-6). *Orthogonal versus oblique factor rotation: A review of the literature regarding the pros and cons* [Paper presentation]. The AAnnual Meeting of the 27th Mid-South Educational Research Association, New Orleans, LA, United States.

Kline, P. (2014). *An easy guide to factor analysis*. Routledge. https://doi.org/10.4324/9781315788135

Kline, T.J. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage publications. https://doi.org/10.4135/9781483385693

Ledesma, R.D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research, and Evaluation*, *12*(1), 2.

Ledesma, R.D., Valero-Mora, P., & Macbeth, G. (2015). The scree test and the number of factors: a dynamic graphics approach. *The Spanish Journal of Psychology*, *18*, Article e11.

Lei, W., Wang, X., Dai, D.Y., Guo, X., Xiang, S., & Hu, W. (2022). Academic self-efficacy and academic performance among high school students: A moderated mediation model of academic buoyancy and social support. *Psychology in the Schools*, *59*(5), 885-899. https://doi.org/10.1002/pits.22653

Li, C., Dewaele, J.M., & Jiang, G. (2020). The complex relationship between classroom emotions and EFL achievement in China. *Applied Linguistics Review*, *11*(3), 485-510.

Li, C., Jiang, G., & Dewaele, J.-M. (2018). Understanding Chinese high school students' foreign language enjoyment: validation of the Chinese version of the foreign language enjoyment scale. *System*, *76*, 183-196. https://doi.org/10.1016/j.system.2018.06.004

Li, C., Wei, L., & Cai, J. (2023). Life satisfaction and L2 engagement in adolescents. *ELT Journal*, *78*(2), 149-159. https://doi.org/10.1093/elt/ccad054

Lin, Z. (2021). A methodological review of machine learning in applied linguistics. *English Language Teaching*, *14*(1), 74-85.

Luo, T., & Derakhshan, A. (2024). Examining the role of classroom climate and teacher-student relationships in EFL students' perceived learning outcomes: A self-determination theory perspective. *Learning and Motivation*, *88*, Article 102062. https://doi.org/10.1016/j.lmot.2024.102062

McDonald, R.P. (1999). *Test Theory: A Unified Treatment*. Lawrence Erlbaum Associates.

MacIntyre, P.D. (2016). *So far so good: An overview of positive psychology and its contributions to SLA*. Springer International Publishing.

Marsh, H.W., Ellis, L.A., Parada, R.H., Richards, G., & Heubeck, B.G. (2005). A short version of the Self Description Questionnaire II: operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment*, *17*(1), 81. https://doi.org/10.1037/1040-3590.17.1.81

Martin, A.J., Colmar, S.H., Davey, L.A., & Marsh, H.W. (2010). Longitudinal modelling of academic buoyancy and motivation: Do the 5Cs hold up over time? *British Journal of Educational Psychology*, *80*(3), 473-496. https://doi.org/10.1348/000709910x486376

Martin, A.J., & Marsh, H.W. (2006). Academic resilience and its psychological and educational correlates: A construct validity approach. *Psychology in the Schools*, *43*(3), 267-281.

Martin, A.J., & Marsh, H.W. (2008a). Academic buoyancy: Towards an understanding of students' everyday academic resilience. *Journal of School Psychology*, *46*(1), 53-83.

Martin, A.J., & Marsh, H.W. (2008b). Workplace and academic buoyancy: Psychometric assessment and construct validity amongst school personnel and students. *Journal of Psychoeducational Assessment*, *26*(2), 168-184. https://doi.org/10.1177/0734282907313767

Martin, A.J., & Marsh, H.W. (2009). Academic resilience and academic buoyancy: Multidimensional and hierarchical conceptual framing of causes, correlates and cognate constructs. *Oxford Review of Education*, *35*(3), 353-370. https://doi.org/10.1080/03054980902934639

Mercer, S. (2019). Language Learner Engagement: Setting the Scene. In: Gao, X. (eds) *Second Handbook of English Language Teaching*. Springer. https://doi.org/10.1007/978-3-030-02899-2_40

Meyer, J., Fleckenstein, J., Retelsdorf, J., & Köller, O. (2019). The relationship of personality traits and different measures of domain-specific achievement in upper secondary education. *Learning and Individual Differences*, *69*, 45-59. https://doi.org/10.1016/j.lindif.2018.11.005

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315-341. https://doi.org/10.1007/s10648-006-9029-9

Putwain, D.W., & Daly, A.L. (2013). Do clusters of test anxiety and academic buoyancy differentially predict academic performance? *Learning and Individual Differences*, *27*, 157-162.

Putwain, D.W., Daly, A.L., Chamberlain, S., & Sadreddini, S. (2015). Academically buoyant students are less anxious about and perform better in high-stakes examinations. *British Journal of Educational Psychology*, *85*(3), 247-263. https://doi.org/10.1111/bjep.12068

Ren, Y., Ai, L., Yang, Y., & Zhu, F. (2025). Examining the influence of aerobic exercise on middle school students' academic engagement: exploring the role of academic buoyancy. *International Journal of Sport and Exercise Psychology*, 1-22. https://doi.org/10.1080/1612197X.2024.2437483

Ryan, R.M., & Deci, E.L. (2024). Self-determination theory. In *Encyclopedia of quality of life and well-being research* (pp. 6229-6235). Cham: Springer International Publishing.

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *NPJ Schizophrenia*, *5*(1), 9. https://doi.org/10.1038/s41537-019-0077-9

Rudd, G., Meissel, K., & Meyer, F. (2021). Measuring academic resilience in quantitative research: A systematic review of the literature. *Educational Research Review*, *34*, Article 100402. https://doi.org/10.1016/j.edurev.2021.100402

Schmidt, A.F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, *98*, 146-151. https://doi.org/10.1016/j.jclinepi.2017.12.006

Seligman, M.E., & Csikszentmihalyi, M. (2014). Positive psychology: An introduction. In *Flow and the foundations of positive psychology: The collected works of Mihaly Csikszentmihalyi* (pp. 279-298). Springer Netherlands. https://doi.org/10.1007/978-94-017-9088-8_18

Shatte, A.B.R., Hutchinson, D.M. & Teague, S.J. 2019. Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(9). 1426-1448.

Sudina, E., & Plonsky, L. (2021). Academic perseverance in foreign language learning: An investigation of language-specific grit and its conceptual correlates. *The Modern Language Journal*, *105*(4), 829-857. https://doi.org/10.1111/modl.12738

Teimouri, Y., Plonsky, L., & Tabandeh, F. (2020). L2 grit: Passion and perseverance for second-language learning. *Language Teaching Research*, *26*(5), 893-918.

Vrdoljak, G., Kurtović, A., & Lovaković, I. (2018). Personality traits, goal orientations and school achievement. *Primenjena Psihologija, 11*(3), 325–344. https://doi.org/10.19090/pp.2018.3.325-344

Wiboolyasarin, K., Kamonsawad, R., Jinowat, N., & Wiboolyasarin, W. (2022). EFL learners' preference for corrective feedback strategies in relation to their self-perceived levels of proficiency. *English Language Teaching Educational Journal*, *5*(1), 32-47.

Worthington, R.L., & Whittaker, T.A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806-838. https://doi.org/10.1177/0011000006288127

Yang, S., Azari Noughabi, M., & Jahedizadeh, S. (2024). Modelling the contribution of English language learners' academic buoyancy and self-efficacy to L2 grit: Evidence from Iran and China. *Journal of Multilingual and Multicultural Development*, *45*(7), 2701-2717. https://doi.org/10.1080/01434632.2022.2062368

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122. https://doi.org/10.1177/1745691617693393

Yao, Y., Wang, W., & Yang, X. (2021). Perceptions of the inclusion of automatic writing evaluation in peer assessment on EFL writers' language mindsets and motivation: A short-term longitudinal study. *Assessing Writing*, *50*, Article 100568. https://doi.org/10.1016/j.asw.2021.100568

Yun, S., Hiver, P., & Al-Hoorie, A.H. (2018). Academic buoyancy: exploring learners' everyday resilience in the language classroom. *Studies in Second Language Acquisition*, *40*(4), 805-830. https://doi.org/10.1017/S0272263118000037

APPENDICES

Appendix 1. *Interview questions in Phase 1 (Original Chinese).*

1. 您能分享一下您从小到大在学习英语过程中遇到的日常挫折是什么吗？这些挫折对您产生了哪些影响？

2. 面对这些挫折，您采取了哪些具体的策略来克服困难并提升英语水平？这些策略对您来说效果如何？

3. 在学习英语的过程中，您如何看待个人的作用？您觉得个人因素在应对学习挑战中扮演了什么样的角色？

4. 您在学习英语的不同阶段（如童年、青年、成年）中，面对挫折时的心态和应对方式是否有变化？如果有，请详细描述这些变化。

5. 您认为除了个人努力外，外部因素（如环境、教师、同伴等）在帮助您克服学习英语的挫折中起到了哪些作用？

6. 您有没有通过某些特定的习惯或方法来保持并持续提高您的英语水平？这些方法是否有效地帮助您适应新的学习环境或提高语言能力？

English version (translated):

1. Can you share what daily frustrations you have encountered in learning English since childhood? What impact have these frustrations had on you?

2. In the face of these frustrations, what specific strategies have you adopted to overcome difficulties and improve your English level? How effective are these strategies for you?

3. How do you view the role of the individual in the process of learning English? What role do you think personal factors play in coping with learning challenges?

4. Has your mentality and coping style changed when facing frustrations in different stages of learning English (such as childhood, youth, and adulthood)? If so, please describe these changes in detail.

5. In addition to personal efforts, what role do you think external factors (such as environment, teachers, peers, etc.) play in helping you overcome frustrations in learning English?

6. Do you have any specific habits or methods to maintain and continuously improve your English level? Are these methods effective in helping you adapt to the new learning environment or improve your language skills?

Appendix 2. *Rating rubric in inter-rater agreement assessment in Phase 1.*

Likert scale (e.g., 1 to 4, where 1 = strongly disagree and 4 = strongly agree)

1. Relevance: Judge how well each item reflects the construct being measured.

2. Clarity: Evaluate whether the item is clearly written, unambiguous, and uses appropriate language.

3. Comprehensiveness: Determine whether the full set of items covers all key dimensions of the construct.

4. Redundancy*: Identify items that repeat the same idea or are too similar to others, adding length without value.

5. Difficulty*: Assess whether the item's complexity or difficulty level is suitable for the intended population.

Note. * reverse scoring

Appendix 3. *Initial L2BS items in Phase 1.*

Likert scale (e.g., 1 to 4, where 1 = strongly disagree and 4 = strongly agree)

1. 我善于应对英语学习挫折（例如，英语成绩不好、英语学习上的负面评价）。

2. 我不会让英语学习压力压垮我。

3. 我擅长应对英语学习压力。

4. 令我不满意的英语成绩不会影响我的信心。

English version (translated):

1. I'm good at dealing with setbacks in learning English (e.g., bad English mark, negative feedback on my English learning).

2. I don't let English study stress get on top of me.

3. I'm good at dealing with English learning pressures.

4. I don't let an unsatisfactory English score affect my confidence.

Appendix 4. *Hyperparameter configurations for predictive models.*

Model	Hyperparameter
Random Forest	n_estimators = 100; min_samples_split = 2 min_samples_leaf=1; max_features='auto'
XGBoost	n_estimators = 100; max_depth = 6 learning_rate = 0.3; min_child_weight=1 subsample=1; colsample_bytree=1
LightGBM	n_estimators = 100; max_depth = -1 learning_rate = 0.1; num_leaves = 31 min_child_samples=20; subsample=1.0, colsample_bytree=1.0

Appendix 5. *The final version of L2BS.*

	非常不同意	不同意	有点不同意	不知道	有点同意	同意	非常同意
Items	1	2	3	4	5	6	7
1.我善于应对英语学习挫折（例如，英语成绩不好、英语学习上的负面评价）。							
2.我不会让英语学习压力压垮我。							
3.我擅长应对英语学习压力。							
4.令我不满意的英语成绩不会影响我的信心。							

English version (translated):

	Strongly disagree	Disagree	Somewhat disagree	I don’t know	Somewhat agree	Agree	Strongly agree
Items	1	2	3	4	5	6	7
1. I’m good at dealing with setbacks in learning English (e.g., bad English mark, negative feedback on my English learning).							
2. I don’t let English study stress get on top of me.							
3. I’m good at dealing with English learning pressures.							
4. I don’t let an unsatisfactory English score affect my confidence.							