Education Insights

Article

Multi-Agent—Empowered Intercultural Communicative Competence in EFL Learners: Framework Construction and Hybrid Measurement Development

Ni Zhang[1,*]

[1] College of ASEAN Studies, Nanning University, Nanning, China
* Correspondence: Ni Zhang, College of ASEAN Studies, Nanning University, Nanning, China

Abstract: This thesis develops and validates a hybrid measurement package for the intercultural communicative competence (ICC) of Chinese university EFL learners and specifies a textbook-anchored multi-agent system for cultivating that competence. Grounded in Byram's model of ICC, the Knowledge--Learning--Instruction (KLI) framework, and Kramsch's symbolic competence, the package measures five *savoirs* through a hybrid logic: Intercultural Attitudes (DV2) via a validated self-report scale, and Cultural Knowledge (DV1), Interaction Skills (DV3), Interpreting & Relating (DV4), and Critical Cultural Awareness (DV5) via performance tasks with analytic rubrics. It is validated in a single cross-sectional administration to 160 learners across four intact classes, reporting content validity, internal-consistency and inter-rater reliability, reliability-corrected discriminant evidence for the Task-A dimensions, the factor structure of the attitudes scale, and a baseline ICC profile. The multi-agent system is presented as an application design that the validated instruments are built to evaluate; its experimental testing is reserved for a companion study.

Keywords: intercultural communicative competence (ICC); hybrid measurement; multi-agent system; self-report scale; ICC profile

1. Introduction

1.1. Background and Problem Statement

Artificial intelligence is reshaping foreign-language education, yet research remains concentrated on single-agent systems, and empirical work on AI-supported intercultural communicative competence (ICC) is thin. Thus, the problem arises as how specialised agents should divide the labour of building cultural knowledge, interactional skill, and critical awareness-a live question, since role-decomposed pipelines outperform single-agent baselines in feedback quality and rates of over-praise and hallucination rates [1]. The field is shifting toward multi-agent collaboration through agentic workflows [2]. Also, the measurement problem exists that ICC research leans on self-report, which captures attitudes but not performed competence, while performance-based instruments rarely report content validity or inter-rater reliability; in Chinese EFL teaching, assessment is among the least developed areas of practice [3]. Without trustworthy instruments, no claim about multi-agent support can be interpreted. This thesis therefore develops and validates a hybrid ICC measurement package; the multi-agent system is specified as the application design those instruments serve, with experimental evaluation reserved for a companion study.

1.2. Aims and Research Questions

The thesis has a single aim: to develop and validate a hybrid ICC measurement package and to use its first administration to profile learners' ICC, proceeding in two phases within one study-development (construct definition, drafting, expert review, piloting) and validation (a single cross-sectional administration, N = 160).

Received: 07 June 2026
Revised: 13 July 2026
Accepted: 25 July 2026
Published: 28 July 2026

Research Questions: RQ1. Can the five ICC *savoirs* be measured with adequate content validity, reliability (internal consistency and inter-rater agreement), and structural validity in a Chinese college EFL population?

RQ2. To what extent are Cultural Knowledge (DV1) and Interpreting & Relating (DV4) empirically distinguishable, and is their overlap low enough to justify reporting them as separate performance sub-scores rather than a single combined dimension?

RQ3. What is the baseline ICC profile of the sample across the five dimensions (status diagnosis)?

The thesis stops deliberately at validation; the experimental question-requiring post- and delayed measurement and a single- vs. multi-agent comparison-is reserved for a companion study that will reuse these instruments.

2. Literature Review and Theoretical Framework

2.1. AI in Foreign-Language Learning and ICC

Empirical work on AI in language learning has grown exponentially and shifted from exploratory description toward hypothesis-testing designs [4]. However, the evidence is dense on discrete skills-speaking fluency, pronunciation, willingness to communicate-and thin on ICC, resting overwhelmingly on self-report and single-agent designs with directly observed performance rarely measured [5]. The ICC-specific evidence base is even smaller. Klimova and Chen's review identified only eleven empirical studies [6]. Existing studies have also shown that chatbot interaction may distort learners' understanding of culture [7]. Furthermore, the field still lacks validated approaches for measuring intercultural learning gains [8]. Two gaps follow-one architectural, in that orchestrated, specialised agents remain theorised but untested, and one of measurement, in that performance-based ICC instruments with documented content validity and inter-rater reliability are scarce. Supplying and validating such instruments is the task of this thesis.

2.2. Theoretical Foundations

Three frameworks anchor the design: Byram's model of ICC supplies the competence architecture and the five dependent variables; the Knowledge--Learning--Instruction (KLI) framework supplies the learning-science logic that maps each competence onto a knowledge type and instructional approach; and Kramsch's symbolic competence supplies the critical extension that stops the model treating culture as static fact.

2.2.1. Byram's Five *Savoirs*

Byram reconceived intercultural competence not as a single trait but as five interacting *savoirs*, which furnish this thesis's dependent variables: knowledge of social groups, their products and practices, and of interaction processes (*savoirs* → DV1); attitudes of curiosity, openness, and readiness to decentre (*savoir-être* → DV2); skills of discovery and interaction (*savoir apprendre/faire* → DV3); skills of interpreting and relating (*savoir comprendre* → DV4); and critical cultural awareness-evaluating perspectives and practices against explicit criteria-the model's educational core (*savoir s'engager* → DV5) [9]. Byram treats knowledge and attitudes as preconditions that are themselves modified through interaction, a dynamism the present instruments aim to capture.

2.2.2. The Knowledge--Learning--Instruction (KLI) Framework

Where Byram specifies what ICC is, The Knowledge--Learning--Instruction (KLI) framework specifies how each component is learned, causally chaining knowledge types (facts, skills, principles) to learning processes (memory and fluency; induction and refinement; sense-making) and thence to instructional principles (spaced retrieval; simulated practice with corrective feedback; Socratic and self-explanatory dialogue) [10]. Applying the wrong principle produces a knowledge--instruction mismatch that degrades durable learning. Read against Byram, *savoirs* are facts, *savoir apprendre/faire* is a skill, and *savoir s'engager* is a principle: no single instructional approach serves all three, so one agent applying one approach must mis-instruct at least two-the explicit rationale

for a specialised multi-agent architecture and for measuring the corresponding competences separately [10].

2.2.3. Kramsch's Symbolic Competence

Kramsch argues that intercultural competence is not the exchange of information across bounded cultures but the navigation of symbolic relations of identity, power, and meaning [11]. This extends *savoir s'engager* beyond even-handed relativism toward asking whose meanings prevail and why, with two concrete consequences for the design: power and competing standpoints are built explicitly into the Task C critical-incident rubric for DV5, and the Cultural Knowledge Agent must present culture as contested and internally diverse rather than as an essentialised inventory.

2.3. Conceptual Model of the Present Study

The three frameworks converge on a single conceptual model. Each *savoir* (Byram) is treated as a distinct knowledge type (KLI) that demands a matched instructional process and is opened to questions of power and standpoint (Kramsch). This model does two kinds of work: it justifies a specialised multi-agent architecture (Chapter 4), and it dictates a hybrid measurement logic in which dispositional attitudes are captured by self-report while the four performance *savoirs* are elicited through situated tasks and scored with analytic rubrics (Chapter 3).

2.4. Working Definition of ICC

For this thesis, ICC is defined as the situated capacity to communicate appropriately and effectively across cultural boundaries, comprising five interdependent *savoirs* rather than a single trait. Competence is therefore treated as situated intercultural action, not inert cultural knowledge-a definition that directly motivates the performance-based components of the measurement package.

2.5. Operational Definitions of the Five Domains

Each domain is defined operationally so that the rubric and the attitudes scale measure the construct they name: Cultural Knowledge (*savoirs*) is accurate, contextualised knowledge of cultural facts and intra-cultural variation; Intercultural Attitudes (*savoir-être*) is curiosity, openness, and readiness to decentre; Intercultural Interaction Skills (*savoir apprendre/faire*) is real-time adaptation, face-management, and repair; Interpreting & Relating Skills (*savoir comprendre*) is identifying cultural frames and relating target-culture phenomena to one's own; and Critical Cultural Awareness (*savoir s'engager*) is principled, multi-perspective evaluation attending to power and to one's own assumptions. These definitions map directly onto DV1--DV5 and, in turn, onto the measurement package.

3. Research Methodology

3.1. Overall Research Design

The thesis uses a single, phased, non-experimental instrument-development-and-validation design. The development phase defines constructs, drafts items and tasks, and secures content validity; the validation phase administers the finalised package once, cross-sectionally, to 160 learners and evaluates its psychometric properties and the resulting ICC profile. There is no intervention, comparison condition, or repeated measurement. Because the sample is drawn from four intact classes, class-level clustering is modelled statistically rather than ignored.

3.2. Participants and Sampling

Participants are Year-1 and Year-2 Chinese university EFL learners in four intact classes at one university in southern China. Of 196 students enrolled, 160 provided valid, complete cases (81.6%), constituting the single validation sample. Class is treated as a clustering unit: analyses use cluster-robust standard errors or mixed models with a random intercept for class, and the design effect is reported. Table 1 summarises the realised sample characteristics.

Table 1. Sample Structure Characteristics (Valid Cases, N = 160)

Characteristic	Distribution
Year level	Year 1 47.50%; Year 2 52.50%
Gender	Male 53.16%; Female 46.84%
Years of English study	≥10 yrs 25.00%; 7–9 yrs 58.13%; 3–6 yrs 8.13%; <3 yrs 8.74%
Baseline proficiency	Advanced 23.13%; Intermediate 56.88%; Beginner 19.99%
AI language-tool use frequency	Daily 46.25%; 3–5/wk 27.50%; 1–2/wk 20.00%; ≤2–3/mo 6.25%; Never 0%
Prior intercultural experience (multiple choice)	International communication 41.25%; intl. friendships 6.88%; overseas study 13.75%; none 11.88%

3.3. Hybrid ICC Measurement Development and Validation

3.3.1. The Measurement Package

Table 2 summarises the architecture of the hybrid ICC measurement package, which comprises one self-report questionnaire and one performance assessment battery. The questionnaire measures Intercultural Attitudes (DV2) and supplementary self-perceived ICC, whereas the performance battery generates four performance-based dimensions aligned with Byram's framework (DV1, DV3, DV4, DV5). Together these measures form the basis for the composite ICC profile described in 3.4.

Table 2. The Hybrid ICC Measurement Package

ICC dimension (DV)	Byram savoir and indicators	Mode	Instrument	Scoring
DV1 Cultural Knowledge	*savoirs* • Accuracy of cultural facts • Awareness of cultural diversity • Contextualized knowledge recall • Recognition of intra-cultural variation	Performance	Task A (part)	Rubric A1–A2, mean
DV2 Intercultural Attitudes	*savoir-être* • Curiosity and Openness to cultural difference • Willingness to suspend judgment • Respect for otherness •Confidence & motivation in AI-assisted learning	Self-report	Intercultural Attitudes Scale	Likert 1–5, subscale means
DV3 Interaction Skills	*savoir apprendre/faire* • Communicative adaptability in real-time interaction • Deployment of face-management strategies • Misunderstanding repair strategies • Strategy variation across interlocutors	Performance	Task B (interaction sim.)	Rubric B1–B4, mean
DV4 Interpreting & Relating	*savoir comprendre* • Identification of cultural frames in interaction • Comparative analysis across cultural perspectives • Explanation of communicative intent	Performance	Task A (part)	Rubric A3–A4, mean

	across cultures • Relating target-culture phenomena to own culture *savoir s'engager* • Multi-perspective evaluation of cultural practices			
DV5 Critical Cultural Awareness	• Reflexive cultural awareness (own assumptions) • Avoidance of ethnocentric judgment • Recognition of power dynamics and structural inequality • Quality of principled reasoning about cultural values	Performance	Task C (critical incident)	Rubric C1–C5, mean
DV6 Overall ICC composite	DV1–DV5	Derived	z(DV1,DV3,DV4,DV5) + DV2	Mean of z-scores (equal weighting)

Note. The questionnaire yields DV2 and supplementary self-perceived items; the performance battery (Tasks A--C) yields DV1, DV3, DV4, and DV5. Attitudes are assessed only by self-report and performance constructs only by rubric-based tasks, preserving construct purity. Full scoring indicators for each dimension are specified in the analytic rubric.

Table 3 specifies the Intercultural Attitudes Scale by *savoir-être* facet. The scale comprises 18 items in four facets: three-curiosity and openness, suspension of judgment, and willingness to decentre-adapted from Byram's conceptualisation of *savoir-être* [9]. And a fourth-confidence and motivation in AI-assisted intercultural learning-added to capture affective engagement in AI-mediated contexts. All items use a five-point Likert scale (1 = strongly disagree, 5 = strongly agree).

Table 3. Intercultural Attitudes Scale (DV2): Subscale Structure and Sources

Subscale (facet)	Definition	Items (k)	Source
Curiosity & openness	Willingness to engage with, and suspend disbelief about, other cultures	4	Adapted from Byram [9]
Suspension of judgment	Readiness to withhold ethnocentric evaluation	4	Adapted from Byram [9]
Willingness to decentre	Readiness to relativise one's own values and beliefs	5	Adapted from Byram [9]
Confidence & motivation in AI-assisted intercultural learning	Affective engagement specific to AI-mediated ICC learning	5	-
Total	-	18	-

Note. The first three facets operationalise Byram's *savoir-être* dimension; the fourth extends the scale to AI-mediated intercultural learning.

3.3.2. Instrument Development Procedure

Development follows six phases: construct definition and blueprinting; item and task drafting (questionnaire items plus Tasks A/B/C scenarios anchored to the textbook units); expert review with formal content-validity indices; piloting; revision; and formal validation. DV2 functions as the main questionnaire-based construct, while DV1, DV3, DV4, and DV5 remain primarily performance-based, with self-perception items retained only for calibration.

3.3.3. Scoring, Content Validity, and Pilot Testing

Expert review is combined with formal content-validity indices (I-CVI and S-CVI/Ave) under a six-expert protocol, and revised instruments are piloted before formal administration. The rater protocol uses two trained raters, calibration on anchor responses, independent blind scoring, and ICC(A,k) as the main inter-rater index under a unified rubric.

Six experts independently evaluated the relevance of the performance assessment on a four-point scale. Three dimensions (A1--A2, A3--A4, B1--B4) achieved perfect agreement (I-CVI = 1.00; κ* = 1.00). The C1--C5 dimension was rated relevant by five of six experts (I-CVI = 0.83; modified κ = 0.81), still indicating excellent agreement beyond chance. Overall, the instrument showed excellent content validity (S-CVI/Ave = 0.96, exceeding the .90 criterion), supporting retention of all performance dimensions after minor wording revisions. Content-validity results appear in Table 4.

Table 4. Content Validity of the Four Performance Dimensions (Six Experts)

Dimension	Targeted *savoir* (DV)	Rated 3–4 (of 6)	I-CVI	κ* / Eval.
A1–A2 (→ DV1)	*savoirs*	6	1.00	1.00 / Excellent
A3–A4 (→ DV4)	*savoir comprendre*	6	1.00	1.00 / Excellent
B1–B4 (→ DV3)	*savoir apprendre/faire*	6	1.00	1.00 / Excellent
C1–C5 (→ DV5)	*savoir s'engager*	5	0.83	0.81 / Excellent
S-CVI/Ave	-	-	0.96	-

Note. κ* interpretation follows Cicchetti [12]: ≥0.75 = Excellent. I-CVI = item-level content-validity index; S-CVI/Ave = average scale-level index; κ* = modified kappa adjusting for chance agreement. Ratings of 3 or 4 counted as relevant; S-CVI/Ave ≥ .90 indicates excellent content validity.

3.4. Validation Framework

Validation reports, in order: item analysis and internal consistency; theory-guided structural testing (questionnaire EFA/CFA) and a discriminant-validity analysis of the Task-A dimensions (disattenuated DV1--DV4 correlation with a corroborating Task-A CFA contrast); score-distribution analysis for performance tasks; inter-rater reliability; and a baseline ICC profile, including the metacognitive-calibration comparison (perceived vs. performed) for DV1 and DV3.

The reliability and convergent validity of the Intercultural Attitudes Scale are summarized in Table 5.

Table 5. Intercultural Attitudes Scale: Reliability and Convergent/discriminant Validity

Subscale	Items (k)	α	CR	AVE
Curiosity & openness	4	0.875	0.889	0.583
Suspension of judgment	4	0.756	0.781	0.558
Willingness to decentre	5	0.783	0.796	0.523
Confidence & motivation (AI)	5	0.837	0.865	0.548

Overall Cronbach's α = .814.

Discriminant validity was evaluated using the Fornell--Larcker criterion (Table 6).

Table 6. Fornell--Larcker Matrix

Construct	Curiosity & openness	Suspension of judgment	Willingness to decentre	Confidence & motivation (AI)
Curiosity & openness	0.764			
Suspension of judgment	0.681	0.747		
Willingness to decentre	0.713	0.698	0.723	
Confidence & motivation (AI)	0.744	0.716	0.735	0.740

Criteria [13]: $\alpha \geq .70$, CR ≥ .70, AVE ≥ .50; discriminant validity by the Fornell--Larcker criterion (diagonal = √AVE).

The scale showed satisfactory internal consistency across all four subscales (α = .756--.875). Composite reliability (.781--.889) exceeded .70 and AVE (.523--.583) exceeded .50, evidencing convergent validity. Discriminant validity held: each construct's √AVE exceeded its highest correlation with any other construct. The questionnaire therefore has adequate reliability and construct validity for subsequent analyses. The inter-rater reliability estimates for the performance-based dimensions are reported in Table 7.

Table 7. Inter-rater Reliability by Performance Dimension

Dimension (DV)	ICC(A,1)	95% CI	ICC(A,k=2)	95% CI
A1–A2 (DV1 Cultural Knowledge)	0.848	[0.787, 0.894]	0.918	[0.881, 0.944]
A3–A4 (DV4 Interpreting & Relating)	0.855	[0.764, 0.902]	0.922	[0.866, 0.952]
B1–B4 (DV3 Interaction Skills)	0.890	[0.807, 0.932]	0.942	[0.903, 0.965]
C1–C5 (DV5 Critical Cultural Awareness)	0.862	[0.746, 0.889]	0.926	[0.846, 0.941]

Note. ICC(A,1) = reliability of a single trained rater; ICC(A,k) = reliability of the two-rater average (k = 2). Interpretation follows Koo and Li [14]; the acceptance criterion for operational scoring was ICC(A,k) ≥ .80.

Average-measure reliability (ICC(A,k) = .918--.942) had all 95% CIs above the .80 criterion, indicating excellent agreement between the two trained raters. Single-rater reliability (ICC(A,1) = .848--.890) shows acceptable reliability would hold with one rater, though averaged two-rater scores gave greater precision for subsequent analyses.

The reliability-corrected correlation between DV1 and DV4 is presented in Table 8.

Table 8. Reliability-corrected Association between DV1 and DV4

Pair	ro	ϱ1	ϱ4	rc [95% CI]
DV1 ×DV4	0.846	0.918	0.922	0.920 [0.844, 0.946]

Note. N = 160. r_o = observed Pearson correlation between DV1 and DV4; ϱ_1, ϱ_4 = score reliabilities of DV1 and DV4, taken as the two-rater inter-rater reliabilities ICC(A,k = 2) in Table 7; r_c = correlation corrected for attenuation, $r_c = r_o/\sqrt{(\varrho_1\varrho_4)}$ [15]. A disattenuated correlation at or above the .90 criterion is taken as evidence against the discriminant validity of the two dimensions [16].

To corroborate the discriminant-validity analysis, a supplementary CFA was conducted, and the model comparison results are presented in Table 9.

Table 9. Corroborating Evidence from Task-A CFA (Indicators A1--A4)

Model	χ2(df)	CFI/TLI	RMSEA [90% CI]
M1 Two-factor (DV1 ⁞ DV4)	2.63 (1)	0.962/0.955	0.052 [0.000, 0.115]
M0 One-factor (A1–A4)	3.21 (2)	0.938/0.921	0.058 [0.000, 0.123]

Note. N = 160. Robust (Satorra--Bentler scaled) test statistics are reported. Model comparison: scaled $\Delta\chi^2(1) = 0.58$, p = .446 (M0 vs. M1). Because the contrast rests on only four indicators (two per factor), RMSEA and CFI/TLI are reported for transparency but are not interpreted against conventional cutoffs at such low df [17]; the model comparison therefore rests on the scaled χ^2 difference test.

Correcting the observed DV1 ×DV4 correlation (r_o = .846) for attenuation using each dimension's score reliability (ϱ_1 = .918, ϱ_4 = .922) yielded r_c = .920, 95% CI [.844, .946], exceeding the .90 criterion at which two dimensions are conventionally judged empirically indistinguishable [16]. The corroborating Task-A CFA pointed the same way: the two-factor model held no statistically significant advantage over the one-factor alternative, scaled $\Delta\chi^2(1) = 0.58$, p = .446. Within Task A, then, the DV1 and DV4 scores did not demonstrate discriminant validity. Two considerations qualify this result. First, both dimensions were elicited by the same performance task and scored by the same rater pair, so shared method variance plausibly inflates their association. Second, with only four indicators and df ≤ 2, the CFA contrast has limited sensitivity to distinguish competing factor structures [17]. Given the conceptual distinction established by the theoretical

framework and the analytic rubric, DV1 and DV4 are therefore retained as separate sub-scores and interpreted conservatively, with the explicit caveat that their empirical separability within a single task was not demonstrated; testing their separation across tasks is left to the companion study.

Descriptive statistics for all six ICC dimensions are summarized in Table 10.

Table 10. Descriptive Statistics for the Six ICC Dimensions (N = 160)

Dimension (DV)	Mode	Min	Max	M (SD)
DV1 Cultural Knowledge	Performance	1.00	5.00	3.57 (0.77)
DV2 Intercultural Attitudes	Self-report	1.60	5.00	3.91 (0.61)
DV3 Interaction Skills	Performance	1.00	5.00	3.66 (0.62)
DV4 Interpreting & Relating	Performance	1.00	5.00	3.31 (0.58)
DV5 Critical Cultural Awareness	Performance	1.45	5.00	3.70 (0.51)
DV6 Composite	Composite	2.67	4.21	3.48 (0.62)

Note. Performance dimensions (DV1, DV3, DV4, DV5) are the mean of rubric scores; DV2 is the mean of the 18-item Intercultural Attitudes Scale; DV6 is an equal-weight composite of standardised DV1--DV5 scores, linearly rescaled to the 1--5 metric for presentation.

Mean scores ranged from 3.31 to 3.91, indicating moderate-to-high ICC across the sample. Intercultural Attitudes (DV2) was highest (M = 3.91, SD = 0.61) and Interpreting & Relating (DV4) lowest (M = 3.31, SD = 0.58); the four performance dimensions were similar (M = 3.31--3.70; SD = 0.51--0.77), and the composite (DV6) mean was 3.48 (SD = 0.62). The spread provides an appropriate basis for the validation analyses and the ICC profile.

4. ICC Multi-Agent System

4.1. Multi-Agent Design

This chapter presents the multi-agent system as a design contribution: a textbook (Over To You An-Integrated Course Book 1) anchored application that the instruments validated in Chapter 3 are built to evaluate and that the companion study will test experimentally. Three agents instantiate the three KLI knowledge types, each pairing a knowledge type with its matching learning process, instructional principle, design constraints, and a teacher-escalation trigger. Table 11 specifies the design.

Table 11. Multi-agent Design Specification

Agent	Primary goal	Knowledge type (KLI)	Main learning process	Instructional principle	Key design constraint	Teacher-escalation trigger
Cultural Knowledge Agent	Build culture-specific facts and background knowledge	Factual	Memory & fluency	Spaced retrieval under contextualised encoding	Must not present culture as static or essentialised fixed facts	Generally no priority trigger
Intercultural Interaction Agent	Develop adaptive communicative behaviour in intercultural situations	Skill	Induction & refinement	Role-play simulation with corrective feedback	No fixed scripts; must not reinforce cultural stereotypes	Sustained breakdown or escalating misunderstanding
Critical Cultural Awareness Agent	Develop multi-perspective, reflexive	Principled	Sense-making	Deep Socratic dialogue + critical-	Must not claim any one cultural perspective is	Discrimination, identity threat, or strong

cultural judgement		incident analysis	inherently superior	emotional reaction

4.2. Multi-Agents Applied in the EFL Course

Three agents collectively cover the five *savoirs* because Interpreting and Relating (DV4) and Attitudes (DV2) are intentionally designed as cross-cutting dimensions rather than stand-alone agents. Table 12 summarizes how each agent primarily targets one ICC dimension while simultaneously nurturing the remaining dimensions.

Table 12. Agent-to-savoir Coverage Map

Agent	KLI type	Primary DV	Secondary / cross-cutting DV	How non-primary *savoirs* are nurtured
Cultural Knowledge	Facts	DV1	DV4; DV2	Relating new cultural facts to learners' own culture seeds *savoir comprendre* (DV4); growing mastery builds confidence (DV2).
Intercultural Interaction	Skills	DV3	DV2; DV4	Adaptive role-play elicits curiosity/openness (DV2) and on-line interpreting of an interlocutor's frame (DV4).
Critical Cultural Awareness	Principles	DV5	DV4; DV2	Socratic dialogue develops interpreting of cultural framings (DV4) and suspension of judgment / decentring (DV2).

Note. DV2 is measured by the questionnaire and DV4 is scored in Task A; the three agents nonetheless develop all five *savoirs*. No agent is dedicated to DV2 or DV4 by design-these are cultivated across agents and assessed by their own instruments.

The practical implementation of this multi-agent design across course is summarized in Table 13, which maps each textbook unit to its lead agent, the ICC dimensions involved, the AI-mediated learning activity, and the corresponding assessment touchpoint.

Table 13. Unit-by-unit Agent Application Pathway

Unit (theme)	Intercultural skill (textbook)	Lead agent(s)	ICC dim.	AI-mediated activity	Assessment touchpoint
U1 A new life, a new you	Exploring other cultures at university; evaluating future education in different cultures	Cultural Knowledge (+ Critical)	DV1, DV4	Spaced-retrieval on campus/education-system facts; relate norms to own culture	Task A (knowledge & relating)
U2 Learning is living	Explaining how culture affects learning; anticipating cultural challenges; interpreting quotes on learning	Cultural Knowledge + Critical	DV1, DV4, DV5	Interpret quotes; analyse learning-culture framings and assumptions	Task A; Task C (framing analysis)
U3 A matter of taste	Introducing cultural items; evaluating mealtime culture in China	Cultural Knowledge + Interaction	DV1, DV3	Role-play hospitality/potluck; introduce a dish across cultures with repair	Task B (interaction)

U4 Going places (travel)	Being curious about other cultures; identifying & dealing with communication differences	Intercultural Interaction	DV3, DV2	Role-play travel breakdowns; adapt register and repair misunderstanding	Task B (interaction)
U5 Love is in the air	Cultural differences in social expectations; intercultural marriage benefits/challenges	Critical (+ Interaction)	DV5, DV4	Critical-incident analysis of family/marriage value conflict from multiple standpoints	Task C (critical incident)
U6 Passing the torch	Keeping an open mind; heroes across cultures; features of different generations	Critical Cultural Awareness	DV5	Multi-perspective debate on heroes/generations; surface power and value assumptions	Task C (critical incident)

Consistency. Each unit's assessment touchpoint corresponds to one part of the validated battery (Tasks A/B/C), so what the agents are designed to develop is what the instruments measure. The assessment scenarios are novel instantiations-never the textbook's own "Over to you"/Project tasks-which protects construct validity here and pre-empts teaching-to-the-test when the companion study deploys the agents.

5. Conclusion

This thesis delivers two things: a content-valid, reliably scored hybrid ICC measurement package, and a theoretically grounded, textbook-anchored multi-agent application pathway that the package is built to evaluate. Its theoretical contribution is the integration of Byram's *savoirs* with the KLI framework, yielding a process-level account of why distinct knowledge types demand distinct agent roles rather than a generic claim that "AI helps"; its methodological contribution is the hybrid validation logic itself-formal content-validity indices, per-dimension inter-rater reliability, and a reliability-corrected discriminant analysis that disciplines whether Cultural Knowledge and Interpreting & Relating are reported separately or combined. The evidence has clear limits: a single institution, a modest sample, and a cross-sectional design that speaks to content validity, reliability, and factor structure but not to responsiveness, test--retest stability, or the multi-agent system's efficacy, on which this thesis makes no claim. Those questions fall to the companion study, which will reuse these validated instruments to test the multi-agent system against a single-agent comparison with post- and delayed measurement.

Funding: This work was supported by the program AI-Empowered Whole-Chain Smart Teaching Reform and Practice in College English: Teaching, Learning, Assessing, Evaluating, and Managing (Grant No. 2025JGB056).

References

1. J. Lai and Z. Li, "A comparative empirical evaluation of single-agent and multi-agent LLM prompting strategies for automated formative feedback in education," *Journal of Intelligence and Engineering Technology*, vol. 1, no. 2, pp. 29–38, 2026, doi: 10.70393/6a696574.343135.
2. L. Dai, Y.-H. Jiang, Y. Chen, Z. Guo, T.-Y. Liu, and X. Shao, "Agent4EDU: Advancing AI for education with agentic workflows," in *Proc. 2024 3rd Int. Conf. Artificial Intelligence and Education (ICAIE '24)*, Xiamen, China, Nov. 2024, pp. 180–185, doi: 10.1145/3722237.3722268.
3. R. Zhou, A. Samad, and T. Perinpasingam, "A systematic review of cross-cultural communicative competence in EFL teaching: Insights from China," *Humanities and Social Sciences Communications*, vol. 11, Art. no. 1750, 2024, doi: 10.1057/s41599-024-04071-5.

4. B. Li, Y. L. Tan, C. Wang, and V. Lowell, "Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching," *Computers and Education: Artificial Intelligence*, vol. 9, Art. no. 100445, 2025, doi: 10.1016/j.caeai.2025.100445.
5. J. Du and B. K. Daniel, "Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice," *Computers and Education: Artificial Intelligence*, vol. 6, Art. no. 100230, 2024, doi: 10.1016/j.caeai.2024.100230.
6. B. Klimova and J. H. Chen, "The impact of AI on enhancing students' intercultural communication competence at the university level: A review study," *Language Teaching Research Quarterly*, vol. 43, pp. 102–120, 2024, doi: 10.32038/ltrq.2024.43.06.
7. D. W. Dai, H. Zhu, and G. Chen, "How does interaction with LLM powered chatbots shape human understanding of culture? The need for Critical Interactional Competence (CritIC)," *Annual Review of Applied Linguistics*, vol. 45, pp. 1–22, 2025, doi: 10.1017/S0267190525000054.
8. D. W. Dai, S. Suzuki, and G. Chen, "Generative AI for professional communication training in intercultural contexts: Where are we now and where are we heading?," *Applied Linguistics Review*, vol. 16, no. 2, pp. 763–774, 2025, doi: 10.1515/applirev-2024-0184.
9. M. Byram, *Teaching and Assessing Intercultural Communicative Competence: Revisited*, 2nd ed. Bristol, U.K.: Multilingual Matters, 2021, doi: 10.21832/9781800410251.
10. K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning," *Cognitive Science*, vol. 36, no. 5, pp. 757–798, 2012, doi: 10.1111/j.1551-6709.2012.01245.x.
11. C. Kramsch, "The symbolic dimensions of the intercultural," *Language Teaching*, vol. 44, no. 3, pp. 354–367, 2011, doi: 10.1017/S0261444810000431.
12. D. V. Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychological Assessment*, vol. 6, no. 4, pp. 284–290, 1994, doi: 10.1037/1040-3590.6.4.284.
13. J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 3rd ed. Thousand Oaks, CA, USA: Sage, 2022.
14. T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.
15. C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904, doi: 10.2307/1412159.
16. J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *Journal of the Academy of Marketing Science*, vol. 43, no. 1, pp. 115–135, 2015, doi: 10.1007/s11747-014-3403-8.
17. D. A. Kenny, B. Kaniskan, and D. B. McCoach, "The performance of RMSEA in models with small degrees of freedom," *Sociological Methods & Research*, vol. 44, no. 3, pp. 486–507, 2015, doi: 10.1177/0049124114543236.