

Assessing Linguocultural Competence in Secondary EFL Education: Developing a Four-Dimensional Performance Rubric

Sitorabonu Vokhidova

A. Avloni National Institute of Pedagogical Excellence
Tashkent, Uzbekistan

Abstract. This article develops a prototype analytic rubric for assessing linguocultural competence among pupils in 7th and 8th grades in Uzbekistan. Using a document-based, theory-informed approach, it refines an existing doctoral framework into four assessment dimensions: cognitive, communicative-pragmatic, linguocultural-interpretive, and reflective-axiological competence. Each dimension includes low, medium, and high performance descriptors linked to proposed tasks involving contextual interpretation, role play, cultural explanation, and structured reflection. The rubric distinguishes insufficient evidence from low performance and is designed to support targeted formative feedback. The study focuses on conceptual development and task alignment. It does not provide empirical evidence of reliability, validity, or classroom effectiveness. The resulting framework offers a context-sensitive starting point for assessing linguocultural competence in secondary EFL education. Further research should examine the clarity of the descriptors, learners' responses to the tasks, and agreement between independent assessors before the rubric is used to guide assessment decisions.

Key words: linguocultural competence, intercultural communicative competence, language assessment, analytic rubric, EFL, Uzbekistan.

Introduction. Successful foreign-language communication depends on more than grammatical accuracy and fluency. Learners must also recognise culturally shaped meanings, adapt their language to different social situations, and interpret perspectives that may differ from their own [1]. For English as a foreign language (EFL) learners, these abilities are essential but difficult to assess, especially, knowing facts about another culture does not necessarily mean understanding its meanings, just as speaking English accurately does not guarantee culturally appropriate communication. The central challenge is therefore not simply to determine what learners know, but to understand how they interpret, communicate and reflect in culturally situated contexts.

This study addresses that challenge through the concept of *linguocultural competence*, understood as the ability to connect language with cultural meaning, interpret expressions in context, make appropriate communicative choices and reflect on different perspectives. Although the concept draws on established approaches to intercultural communicative competence, the focus here is practical and assessment-oriented: what can pupils demonstrate when they encounter culturally meaningful situations? To address this question, the study develops a prototype analytic rubric that translates the broad concept of linguocultural competence into observable performance criteria for pupils in grades 7-8 in Uzbekistan.

M.Byram's framework provides an important theoretical foundation through its integration of knowledge, attitudes, skills of interpreting and relating, skills of discovery and interaction, and critical cultural awareness [1]. These elements inform the present rubric in different ways. The cognitive dimension addresses cultural knowledge, the communicative-pragmatic dimension focuses on context-sensitive language use, the linguocultural-interpretive dimension emphasises the interpretation of culturally embedded meanings, and the reflective-axiological dimension focuses on reasoned consideration of perspectives and values. This four-dimensional structure is an assessment-oriented adaptation, not a direct reproduction of Byram's model. The dimensions overlap conceptually, and their empirical distinctness remains to be examined.

Previous research has explored several approaches to intercultural competence assessment. M.Kanik, for example, proposed a discourse-based task for assessing intercultural and sociolinguistic competence [2], while Carpenter and Matsugu examined an analytic rubric for performance-based assessment of intercultural understanding and investigated score interpretation and rater agreement [3]. More recently, A.Serikbayeva examined psychometric instruments for intercultural communicative competence in a Kazakhstani university context [4]. These studies demonstrate that performance-based and psychometric approaches to intercultural assessment already exist. The present study addresses a narrower, context-specific problem: developing a task-linked, four-dimensional rubric intended for pupils in grades 7-8 in Uzbekistan. Its contribution lies in the proposed alignment of cultural knowledge, pragmatic language use, interpretation and reflection within a formative classroom assessment framework, rather than in claiming to introduce an entirely new form of intercultural assessment.

The *Common European Framework of Reference for Languages: Companion Volume* offers a complementary foundation through descriptors for mediation and plurilingual and pluricultural competence [5]. It provides a common reference for articulating learning goals, while the present study focuses on translating selected aspects of intercultural learning into task-specific criteria for pupils in grades 7-8 in Uzbekistan. Neither framework, on its own, establishes whether a particular rubric produces consistent or appropriate assessment judgements in this context.

Research on intercultural assessment points to the importance of systematic instrument development. L.Sercu proposed a framework for developing and evaluating intercultural competence assessments, emphasising the need to consider test quality rather than relying on broad definitions of competence alone [6]. T.Lee studying an online intercultural EFL curriculum with eighth-grade pupils in Thailand, used learner rating scales, Cambridge English examinations and course-evaluation questionnaires. Their study illustrates the contextual adaptation and evaluation of assessment instruments [7], but it did not develop the same four-dimensional performance rubric proposed here.

The broader assessment literature also highlights a need to consider how competence is evidenced. In a systematic review of assessment tools for social, emotional and intercultural competences, F.Müller found that 88.6% of the identified tools relied on self-report. Although their review covered published studies from 2000 to 2017 and addressed several competence areas, it illustrates the value of examining evidence beyond learners' perceptions of their own abilities. Performance tasks may provide a complementary source of evidence by showing how learners interpret contextual cues, choose communicative strategies and explain their reasoning [8].

Analytic rubrics can make assessment criteria more explicit and help organise feedback. A.Jonsson and G.Svingby found that scoring consistency can be improved through analytic, topic-specific rubrics, particularly when supported by exemplars or rater training [9]. However, the use of a rubric does not itself establish the validity of assessment judgements. The *Standards for Educational and Psychological Testing* [10] and M.Kane's argument-based approach to validation both emphasise the evidence needed to support intended score interpretations and uses [11]. For the present study, this means that the descriptors must align with the intended construct and the tasks must elicit evidence relevant to the judgements teachers are expected to make. Fairness also requires that cultural reflection be assessed through reasoned engagement with perspectives, not agreement with an assessor's personal values.

The research gap addressed here is therefore specific. The established frameworks and assessment studies discussed above offer theoretical principles, descriptors or examples of instrument development, but they do not provide the particular task-linked, four-dimensional rubric developed in this study for Uzbek pupils in grades 7-8. This is a context-specific contribution, not a claim that no comparable assessment instrument exists elsewhere. The study develops a prototype comprising cognitive, communicative-pragmatic, linguocultural-interpretive and reflective-axiological dimensions, with observable indicators, three performance levels and proposed tasks for eliciting relevant evidence. It aims to provide a transparent basis for formative assessment while identifying the further evidence needed to evaluate the rubric's clarity, scoring consistency and suitability for classroom use.

Materials and Methods. This study used a document-based, theory-informed approach to develop a prototype analytic rubric for assessing linguocultural competence in secondary English as a foreign language (EFL) education. The purpose was to translate the construct into observable performance criteria and align those criteria with proposed classroom assessment tasks. The intended use is formative: to provide a profile of learners' performance across four dimensions, rather than a definitive measure of general competence. The development process drew on research on intercultural language learning and assessment [12] and on principles of assessment validity and rubric design [10]. In line with these principles, the intended interpretation, target population and proposed uses of the rubric were specified before considering its application. The study concerns the development and conceptual examination of the instrument. It does not report an empirical validation study.

The investigation was conducted during 2024-2025 in seven secondary schools across Tashkent City and the Bukhara and Samarkand regions of Uzbekistan. The study involved 1062 pupils in grades 7-8, including 536 in experimental groups and 526 in control groups, as well as 126 English teachers. The teachers completed questionnaires, and the dissertation documents the original methodological model, assessment criteria and tasks. These details provide the research context from which the present study draws; the participants were not used as an independent validation sample for the revised rubric.

The present study did not reanalyse individual pupil data from the original intervention or generate new reliability estimates. Instead, it examined the documented framework and used it to refine the performance descriptors and develop

task specifications for the proposed rubric. Accordingly, the findings reported here concern the conceptual development and design of the assessment framework, rather than its empirical validation.

Linguocultural competence was defined as the ability to connect linguistic expressions with cultural meanings, interpret those meanings in context, select context-sensitive communicative strategies and reflect on different perspectives. This definition was operationalised through four dimensions:

1. **Cognitive:** identifying and explaining relevant cultural information in context.
2. **Communicative-pragmatic:** choosing language appropriate to the purpose, audience and situation.
3. **Linguocultural-interpretive:** interpreting culturally embedded meanings using contextual evidence.
4. **Reflective-axiological:** examining assumptions, considering alternative perspectives and explaining the basis of an interpretation or evaluation.

The dimensions distinguish related aspects of performance without assuming that they are empirically independent. The rubric assesses observable responses rather than inferring stable personal values from a single task. In particular, reflective performance is judged by the reasoning provided, not by whether a learner agrees with the assessor's cultural viewpoint.

The rubric was developed through five linked stages.

Stage 1: Examination of the original framework. The four components, indicators and low, medium and high-level descriptions in the dissertation were examined to identify the intended ability represented by each criterion. The original wording served as the starting point rather than as a validated scoring scheme for the revised instrument.

Stage 2: Operationalisation of each criterion. Broad descriptions were reformulated in terms of actions or evidence observable in a learner's response. For example, general references to openness were expressed through considering another perspective and explaining whether an initial interpretation should be retained or revised. Similarly, general cultural knowledge was distinguished from the evidence used to interpret a culturally embedded expression.

Stage 3: Differentiation of performance levels. For each dimension, the descriptors were organised around a consistent criterion across three levels. Progression was expressed through differences in contextual accuracy, appropriateness of communicative choices, use of supporting evidence and depth of reasoning, as relevant to the dimension. Low, medium and high were therefore treated as ordered descriptions of task-specific performance, not as interchangeable labels for overall competence.

Stage 4: Mapping tasks to evidence. Four proposed task formats, namely, contextual interpretation, pragmatic role play, cultural explanation to an unfamiliar audience, and structured reflection were linked to the performance evidence they were intended to elicit. The mapping was used to identify whether the task format provided an opportunity to demonstrate the targeted criterion. Table 1 presents this construct-to-evidence map.

Stage 5: Internal coherence review. The prototype was examined against five design considerations: relevance to the specified dimension, observability in a learner response, distinction between adjacent levels, alignment between the task and the intended evidence, and fairness to learners with differing English proficiency and cultural experiences. Particular attention was given to overlap between dimensions,

unsupported cultural generalisations and the risk of treating language errors as evidence of limited cultural understanding when they did not affect the targeted performance.

This procedure produced the prototype descriptors and task specifications presented in the Results section. The internal review was a design analysis, not an independent expert evaluation. No formal expert-rating statistics or pupil pilot results are claimed for the revised version.

The task formats were derived from diagnostic categories documented in the underlying research. Contextual interpretation is intended to elicit recognition of relevant cultural information and an evidence-supported explanation of meaning. Pragmatic role play is intended to elicit context-sensitive requests, refusals or clarification strategies. Cultural explanation is intended to show whether learners can communicate relevant cultural information for an unfamiliar audience. Structured reflection is intended to elicit explanations of assumptions, supporting evidence and alternative perspectives.

Task specifications should identify the situation, addressee, expected response and dimension being assessed. Materials should present cultural practices within a particular context rather than as universal characteristics of a national or linguistic group. Before operational use, the task guide also needs to specify language demands, timing, permitted assistance and administration conditions. These details are necessary for consistent implementation and are proposed requirements for future classroom use, not evidence that a standardised administration procedure has already been validated.

The unit of rating is a dimension demonstrated within a specific task. Each rating is based on the evidence in the learner's response rather than on a general impression of the learner's competence. Before assessment, the task-to-evidence map specifies which dimensions each task is intended to elicit. When a task addresses more than one dimension, each dimension is judged separately and only when the response provides relevant evidence. For each task-by-dimension combination, the assessor assigns one of three performance levels: low, medium or high. **Insufficient evidence (IE)** is recorded when the task is intended to elicit a dimension but the response is missing, incomplete or otherwise provides too little evidence for a judgement. **Not elicited (NE)** is recorded when the task does not provide an opportunity to assess that dimension. Neither category is treated as a performance level. Each rating is accompanied by a brief record of the evidence supporting the judgement and any assistance provided during the task. Linguistic errors are considered in relation to their effect on the specific ability being assessed; they are not automatically treated as evidence of limited cultural understanding.

When the same dimension is assessed through more than one task, the task-specific ratings are retained separately. The prototype does not average these ratings, select a single overall level or calculate a total score. Such aggregation would require explicit rules and supporting evidence. The original doctoral framework allocated up to 25 points to each component, with classifications of 0–55 (low), 56–85 (medium) and 86–100 (high). These historical thresholds are not transferred to the revised rubric. Establishing numerical scores or an overall classification is beyond the scope of the present study.

The analysis was qualitative and design-focused. It examined the correspondence between the construct definition, the four dimensions, the descriptor wording, the proposed tasks and the intended uses of the resulting performance

profile. It did not estimate internal structure, inter-rater agreement, classification accuracy or effects on learning.

The Standards for Educational and Psychological Testing emphasise that assessment interpretations and uses require appropriate supporting evidence [10]. Accordingly, the prototype is presented as a reasoned starting point for classroom assessment, not as a validated measurement instrument. Establishing its quality will require further evidence from expert review, learner response processes, classroom administration and independent scoring by multiple assessors.

Table 1. Task-to-evidence map for the proposed rubric

Nº	Proposed task format	Dimension assessed	Intended evidence
1.	Contextual interpretation	Cognitive	Accurate identification and contextual explanation of relevant cultural information.
2.	Contextual interpretation	Linguocultural-interpretive	Interpretation of culturally embedded meanings using relevant contextual cues.
3.	Pragmatic role play	Communicative-pragmatic	Selection of language appropriate to the purpose, addressee and situation, including clarification or adaptation when needed.
4.	Cultural explanation to an unfamiliar audience	Cognitive	Accurate selection and explanation of relevant cultural information.
5.	Cultural explanation to an unfamiliar audience	Communicative-pragmatic	Presentation of cultural information in language appropriate to the intended audience and communicative purpose.
6.	Structured reflection	Linguocultural-interpretive	Explanation of an interpretation and the contextual evidence supporting it.
7.	Structured reflection	Reflective-axiological	Examination of assumptions, consideration of alternative perspectives and reasoned justification of a position.

Results. The main outcome was a prototype analytic rubric comprising four dimensions, namely, cognitive, communicative-pragmatic, linguocultural-interpretive and reflective-axiological—with three performance levels for each. The framework retains the four dimensions of the original doctoral model but expresses their criteria through observable features of learners' responses. The resulting matrix contains 12 descriptors (Table 2).

The dimensions distinguish four aspects of performance: understanding cultural information, using language appropriately, interpreting culturally embedded meanings and reflecting on different perspectives. These aspects can be recorded separately within a task. However, whether they are empirically distinct in learners' performance remains to be established.

The prototype replaces broad judgements with more specific descriptions of what learners are expected to demonstrate. In the cognitive dimension, the focus is on identifying relevant cultural information and explaining its context. In the communicative-pragmatic dimension, it is on selecting appropriate language for a particular purpose, audience and situation. The linguocultural-interpretive dimension focuses on the evidence learners use to explain cultural meanings, while the reflective-

axiological dimension examines how they justify an interpretation, consider alternative perspectives and reconsider initial assumptions. The latter avoids treating agreement with an assessor's cultural views as evidence of achievement.

The three levels describe progression within each dimension rather than assigning one overall label to a learner's competence. Low performance is characterised by substantial omissions, inappropriate choices or limited justification. Medium performance reflects a relevant but partly developed response. High performance demonstrates the targeted ability through contextual accuracy, appropriate adaptation or reasoned interpretation, as relevant to the dimension. The prototype also distinguishes insufficient evidence from low performance. A missing response or a task that fails to elicit the relevant ability does not provide the same evidence as an attempted response that falls below the criterion. This distinction indicates when further assessment may be needed.

The task-to-evidence map links the four proposed task formats to the dimensions they are intended to assess. Contextual interpretation provides evidence of cultural knowledge and interpretation; pragmatic role play focuses on context-sensitive language use; cultural explanation combines cultural knowledge with audience-appropriate communication; and structured reflection elicits evidence of interpretation and reasoned consideration of perspectives.

The proposed assessment profile records a separate rating for each dimension elicited by a task, accompanied by the supporting response evidence and any assistance provided. If a dimension is assessed through several tasks, the corresponding ratings remain task-specific rather than being combined into a single level. The distinction between insufficient evidence and not elicited also prevents a lack of assessment opportunity from being interpreted as low competence. This procedure makes the intended scoring unit explicit and avoids unsupported aggregation. The results of the present development analysis are the rubric descriptors, the task-to-evidence map and the proposed recording procedure. Whether these procedures lead to consistent judgements in actual classroom use remains to be investigated.

Table 2. Proposed four-dimensional analytic performance rubric

No	Dimension	Low performance	Medium performance	High performance
1.	Cognitive	Identifies little relevant cultural information or gives an inaccurate or unsupported account of its context.	Identifies relevant cultural information accurately but explains its context only partly.	Identifies relevant cultural information accurately and explains its context using evidence from the supplied material.
2.	Communicative-pragmatic	Uses a strategy that does not communicate the intended purpose appropriately for	Communicates the intended purpose through a generally suitable strategy but shows limited	Communicates the intended purpose appropriately for the addressee and situation, adapting or clarifying the

		the addressee or situation.	adaptation or clarification when needed.	response when needed.
3.	Linguocultural-interpretive	Offers a predominantly literal or unsupported interpretation despite relevant contextual cues.	Offers a plausible interpretation supported by a relevant contextual cue but provides limited explanation of the meaning.	Justifies the interpretation with specific contextual evidence and considers alternative meanings where the material permits ambiguity.
4.	Reflective-axiological	Gives an evaluation without explaining its basis or considering another perspective.	Explains a position and acknowledges another perspective but examines initial assumptions only partly.	Examines initial assumptions, considers another perspective and uses reasoned justification to support or revise the interpretation.

Table 3. Refinements from the original criteria to the prototype

No	Dimension	Original assessment emphasis	Refinement in the prototype
1.	Cognitive	Cultural knowledge and general ability to explain a text.	Identifies and explains relevant cultural information in context.
2.	Communicative-pragmatic	Appropriate language use, including independent strategy selection and performance with teacher support.	Specifies purpose, audience, situation, adaptation and clarification; records assistance.
3.	Linguocultural-interpretive	Understanding realia, idioms, metaphors and cultural contexts.	Requires contextual evidence to support interpretations and considers alternative meanings where appropriate.
4.	Reflective-axiological	Tolerance, respect and openness expressed through	Assesses reasoned reflection on assumptions and perspectives rather

		broad evaluative descriptions.	than presumed personal attitudes.
--	--	--------------------------------	-----------------------------------

Discussion. This study developed a prototype four-dimensional analytic rubric for assessing linguocultural competence among pupils in Grades 7-8 in Uzbekistan. Its main contribution is to translate a broad educational aim into observable criteria covering cultural knowledge, context-sensitive language use, interpretation of culturally embedded meanings and reflection on different perspectives. Rather than relying on general impressions of cultural awareness or English proficiency, the framework specifies what learners are expected to demonstrate and what evidence teachers should consider.

An important feature of the rubric is its distinction between cultural knowledge and cultural interpretation. Recognising a cultural reference does not necessarily mean understanding its significance in a particular situation. The cognitive dimension therefore focuses on identifying and explaining relevant cultural information, while the linguocultural-interpretive dimension assesses how learners use contextual evidence to explain meaning. This distinction is consistent with A.Liddicoat and A.Scarino's view of language learning as inseparable from its social and cultural context [12]. In classroom practice, it could help teachers identify whether a learner needs to develop cultural knowledge, interpretive reasoning or both.

The communicative-pragmatic dimension extends assessment from understanding to action. Learners must choose language appropriate to their purpose, addressee and situation, and clarify or adapt their response when necessary. This emphasis complements the *Common European Framework of Reference for Languages: Companion Volume*, which includes descriptors for mediation and plurilingual and pluricultural competence [5]. However, a learner's performance must also be interpreted in light of their English proficiency and the assistance provided. Otherwise, language limitations may be mistaken for limited cultural understanding, or supported performance may be interpreted as independent achievement.

The reflective-axiological dimension addresses a different challenge: assessing reflection without making assumptions about learners' personal values. The original framework used broad concepts such as tolerance, respect and openness. The prototype instead focuses on observable reasoning, including explaining an interpretation, considering another perspective and justifying whether an initial view should be retained or revised. This approach is consistent with A.Scarino's emphasis on assessing both communicative performance and learners' explanations of their thinking [13]. It also avoids treating agreement with an assessor's cultural preferences as evidence of competence.

The three performance levels provide an initial way to describe progression within each dimension. For example, interpretive performance develops from a predominantly literal or unsupported explanation towards one justified by contextual evidence. Pragmatic performance develops through more appropriate language choices and greater capacity to adapt or clarify communication. In reflection, progression concerns the quality of reasoning and consideration of alternative perspectives. Separating these judgements may offer more specific formative feedback than a single overall rating. The additional category of insufficient evidence also helps

distinguish a genuinely limited response from a task that has not provided enough evidence for judgement.

These design choices are consistent with wider research on assessment rubrics. A.Jonsson and G.Svingby found that analytic, topic-specific rubrics, particularly when supported by exemplars and rater training, can improve scoring reliability [9]. E.Panadero reported an overall moderate positive effect of rubrics on academic performance, although their findings do not establish the effectiveness of this particular instrument [14]. The present study contributes at an earlier stage: it defines the intended dimensions, proposes performance descriptors and links them to suitable task formats. It does not yet show that different teachers would rate the same response consistently or that the four dimensions are empirically distinct.

This distinction is important when interpreting the study's evidence. The underlying doctoral investigation involved 1062 pupils and 126 English teachers across seven secondary schools in Uzbekistan. These participants formed part of the broader investigation, not a sample independently assessed with the refined rubric reported here. The intervention and its pretest–posttest results therefore cannot establish the reliability, validity or feasibility of the prototype. The findings of this article concern the development and conceptual alignment of the rubric, and its proposed classroom applications should be treated as intended uses rather than demonstrated effects.

The framework offers several practical possibilities for teachers. A four-dimensional profile could help identify whether a learner needs support with cultural knowledge, pragmatic language choices, interpretation or reflection. Contextual interpretation, role play, cultural explanation and structured reflection could elicit different kinds of evidence and provide a basis for more targeted feedback. For these tasks to be fair, however, cultural content must be carefully contextualised and should not present national cultures as uniform. Assessment should reward evidence-based interpretation and reasoned consideration of perspectives, rather than unsupported generalisations or agreement with a preferred cultural viewpoint.

Several limitations remain. The refined rubric has not undergone empirical validation, and no estimates of inter-rater agreement, classification accuracy or classroom feasibility are available for this version. The proposed tasks also need to be tested to determine whether they elicit adequate evidence for each dimension and whether assessors can distinguish adjacent performance levels consistently. In addition, the framework was designed for pupils in Grades 7-8 in Uzbekistan; its suitability for other age groups, proficiency levels and educational contexts requires further investigation.

Future research should begin with expert review of the descriptors and tasks, followed by interviews with pupils to establish how they understand the instructions and interpret the assessment demands. Classroom piloting should then collect actual learner responses and have multiple assessors rate them independently. Such data would allow researchers to examine agreement between raters, identify ambiguous descriptors and investigate whether the four dimensions are distinguishable in practice. Numerical scoring and classification thresholds should be introduced only after explicit scoring rules have been established and supported by evidence.

In conclusion, this study presents a context-sensitive prototype for assessing linguocultural competence through four proposed dimensions of performance. Its contribution is to connect the aims of intercultural language education with observable criteria and proposed classroom tasks. The rubric offers a structured starting point for

formative assessment, but its practical value will depend on further evidence that the tasks elicit the intended performance and that teachers can apply the descriptors consistently.

References:

1. Byram, M. (2021). *Teaching and assessing intercultural communicative competence: Revisited* (2nd ed.). Multilingual Matters. <https://doi.org/10.21832/9781800410251>
2. Kanik, M. (2013). Reverse discourse completion task as an assessment tool for intercultural competence. *Studies in Second Language Learning and Teaching*, 3(4), 621–644. <https://doi.org/10.14746/sslt.2013.3.4.8>
3. Carpenter, J., & Matsugu, S. (2023). Validating a performance assessment of intercultural understanding. *JLTA Journal*, 26, 23–42. https://doi.org/10.20622/jltajournal.26.0_23
4. Serikbayeva, A., Akizhanova, D., Shinetova, L., Altybayeva, S., & Abdrasilov, B. (2026). Psychometric tools for assessing intercultural communicative competence: A case study for evaluating ICC in Kazakhstan. *Frontiers in Education*, 11, Article 1727766. <https://doi.org/10.3389/feduc.2026.1727766>
5. Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing. <https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4>
6. Sercu, L. (2004). Assessing intercultural competence: A framework for systematic test development in foreign language education and beyond. *Intercultural Education*, 15(1), 73–89. <https://doi.org/10.1080/1467598042000190004>
7. Lee, T.-Y., Ho, Y.-C., & Chen, C.-H. (2023). Integrating intercultural communicative competence into an online EFL classroom: An empirical study of a secondary school in Thailand. *Asian-Pacific Journal of Second and Foreign Language Education*, 8, Article 4. <https://doi.org/10.1186/s40862-022-00174-1>
8. Müller, F., Denk, A., Lubaway, E., Sälzer, C., Kozina, A., Vršnik Perše, T., Rasmusson, M., Jugović, I., Lund Nielsen, B., Rozman, M., Ojsteršek, A., & Jurko, S. (2020). Assessing social, emotional, and intercultural competences of students and school staff: A systematic literature review. *Educational Research Review*, 29, Article 100304. <https://doi.org/10.1016/j.edurev.2019.100304>
9. Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. <https://doi.org/10.1016/j.edurev.2007.05.002>
10. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
11. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
12. Liddicoat, A. J., & Scarino, A. (2013). *Intercultural language teaching and learning*. Wiley-Blackwell. <https://doi.org/10.1002/9781118482070>
13. Scarino, A. (2009). Assessing intercultural capability in learning languages: Some issues and considerations. *Language Teaching*, 42(1), 67–80. <https://doi.org/10.1017/S0261444808005417>
14. Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A

meta-analytic review. *Educational Psychology Review*, 35, Article 113.
<https://doi.org/10.1007/s10648-023-09823-4>