International Journal of TESOL Studies (2026)
260803, 1-21 https://doi.org/10.58304/ijts.260803

Article

Policy–Test Alignment in High-Stakes Assessment: Evaluating Construct Representation in China's NMET Under the 2017 Curriculum Standards

Lu Geng Sun
Jirada Wudthayagorn*
Chulalongkorn University, Thailand

Received: 1 May 2026 / Received in revised form: 29 June 2026 / Accepted: 24 July 2026 / Available online: 9 September 2026

Abstract
This study examines policy–test alignment in *China's National Matriculation English Test* (NMET) by evaluating how assessment evidence represents the competencies articulated in the *Senior High School English Curriculum Standards* (ECS 2017). Recent curriculum reforms emphasize competency-oriented English education, yet the extent to which these policy goals are reflected in high-stakes examinations remains unclear. The analysis uses Kane's Interpretation/Use Argument (IUA) to examine construct representation across NMET task domains, rather than treating alignment as surface correspondence between policy statements and test formats. The documentary analysis focused on ECS 2017 and January and June 2025 NMET papers used in Zhejiang Province, where nationally standardized papers are locally administered. Alignment was examined through expert judgment involving five specialists ($\kappa = .366$) and Webb's alignment indices. The findings indicate that language competence is systematically represented across NMET sections, providing stronger construct representation evidence within this domain. In contrast, cultural awareness and thinking quality are represented selectively and unevenly, which limits the claims that can be made from NMET evidence. Learning ability is not instantiated in NMET task demands. Analysis of construct representation suggests a coherent assessment design but does not offset gaps in competency coverage. This study concludes that NMET evidence should be interpreted in domain-specific terms. It provides stronger evidence for some ECS 2017 competencies than for others, and it cannot support broad claims about the full curriculum framework.

Keywords
Policy-test alignment, Interpretation/Use Argument (IUA), ECS 2017, NMET, construct representation, high-stakes large-scale assessment

*Corresponding author. Email: jirada.w@chula.ac.th

1 Introduction

In recent decades, China's English education policies have increasingly emphasized the development of communicative competence and broader language-related competencies aligned with national goals of modernization and global engagement (Gao, 2018; Gil, 2016; Saeedi et al., 2025). This orientation is formally articulated in the *Senior High School English Curriculum Standards* (ECS 2017), which reconceptualize English learning beyond discrete linguistic knowledge toward integrated language use, cultural awareness, and higher-order thinking (Pan & Wang, 2024). Although ECS 2017 includes speaking within its competence-oriented view of English learning, the incorporation of speaking into high-stakes English tests remains uneven across provincial contexts, with speaking components contributing to the English score in some regions but remaining outside the main scored components in many others (Xu et al., 2025).

The *National Matriculation English Test* (NMET) plays a central role in this policy context because it links senior high school English learning to university admission. As a compulsory component of the *Gaokao* (National College Entrance Examination, NCEE), NMET performance carries decisive consequences for university admission and strongly influences what schools and students prioritize in English learning (Cheng & Hamid, 2025). In the Zhejiang NMET papers analyzed in this study, speaking is not included in the primary scored English components. For this reason, the examination is one of the main places where curriculum goals become visible in assessment practice.

This relationship between curriculum goals and test design raises an important question. Recent curriculum reforms promote competency-oriented English education, but high-stakes examinations operate under practical constraints. Their tasks need to be administered, scored, and compared under common conditions. Competencies related to communicative and integrated language use can therefore be difficult to represent through task formats that permit reliable and efficient scoring (Pan & Qian, 2017; Zhang, 2024). As a result, tensions may arise between curriculum aspirations and the competencies that high-stakes tests are able to prioritize in practice.

This issue is particularly consequential in China, where the NMET functions as a de facto curriculum signal for senior high schools (Cheng, 2008; Cheng & Hamid, 2025). At the provincial level, where national assessment policies are enacted through locally administered test papers, this signaling function becomes especially salient.

Previous NMET research has mainly examined washback, instructional response, and learner experience, while less attention has been given to how the examination itself represents the competency goals of the ECS 2017 through task design. This study addresses this gap by examining construct representation in Zhejiang's NMET from a validity-oriented perspective, with particular attention to how ECS 2017 competencies are represented across test sections and task types.

2 Literature Review

2.1 Policy-test alignment from a validity perspective

Research on language education policy and assessment has paid increasing attention to the relationship between curriculum intentions and high-stakes examinations, particularly in contexts where test outcomes play a central role in shaping instructional priorities and policy enactment (Cheng, 2008; Wallace & Ke, 2023). Within this body of work, the concept of policy-test alignment has been used to capture the extent to which assessment practices are understood to reflect, mediate, or constrain curriculum-based educational goals (Pan & Qian, 2017). In this study, the issue is not only whether the NMET includes similar content to ECS 2017, but whether its tasks provide evidence for interpreting curriculum-defined competencies. This concern is also relevant to language assessment research on standards-assessment relationships, where alignment between language standards frameworks and large-scale assessments is discussed as validity evidence for interpreting assessment results in relation to standards (Gottlieb & Chapman, 2022).

From a validity-oriented perspective, validity concerns the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores (Messick, 1989, p. 13). Kane (2013, p. 1) similarly argues that validating an interpretation or use of test scores involves evaluating the plausibility of the claims based on those scores. For policy-test alignment, this perspective shifts the focus from simple content correspondence to whether test tasks provide evidence for the representation of curriculum-defined competencies in assessment design. This validity concern is especially relevant when high-stakes language assessments are shaped by curriculum policy. As Chalhoub-Deville (2016, p. 454) argues, reform-based testing involves claims that extend beyond individual score interpretation and require attention to the relationship among policy stipulations, testing capabilities, and those affected by assessment practice.

The Interpretation/Use Argument (IUA), developed within Kane's argument-based approach to validation, provides a principled validity framework for evaluating whether proposed interpretations and uses of assessment performance are supported by explicit inferences and evidence (Kane, 2006, 2013). In language assessment research, this approach has been used to organize evidence for score interpretation and use around specified inferences and their supporting assumptions (Chapelle, 2008; Hsieh, 2024). Within the IUA, observed performances are connected to score-based interpretations and uses through inferential links. Scoring links observed performances to observed scores, while generalization extends those scores beyond the immediate testing occasion to broader claims about performance across relevant conditions (Kane, 2013, p. 10). The argument may then involve extrapolation to broader performance domains and, where relevant, decisions or uses based on score interpretations (Kane, 2013, p. 11). These inferential links require warrants and backing, which provide the rationale and evidence for moving from one claim to the next (Kane, 2013, p. 12). Following Chapelle et al. (2008), this study distinguishes construct-focused evidence from extrapolation to performance beyond the test. In this study, construct representation is the immediate analytical focus. The analysis does not directly test extrapolation through evidence of students' performance beyond the examination; instead, it examines whether NMET design, content, and task structure provide a representational basis for claims about what NMET performance can and cannot be taken to represent.

Policy alignment is therefore used as an analytical lens, not as an assumption that assessment design should directly reproduce curriculum policy. The analysis is more limited in scope: it identifies what claims can reasonably be made about the representation of curriculum-defined competencies in the NMET and where the limits of those claims lie.

2.2 The ECS 2017 as a policy construct framework

The ECS 2017 represents a major reorientation of China's English education policy toward a competency-based framework that foregrounds the development of students' comprehensive language ability rather than the accumulation of discrete linguistic knowledge (Pan & Wang, 2024; Tang et al., 2025). The ECS 2017 expresses this shift through a model of English subject core competencies consisting of four domains, rather than through a narrow focus on grammatical knowledge or isolated language skills (MoE, 2020, p. 4). Scholarly interpretations of the ECS 2017 suggest that this reorientation reflects a response to the limitations of exam-driven instruction and an effort to realign English education with communicative and developmental goals in a globalizing context (Pan & Wang, 2024; Wang & Luo, 2019).

At the core of the ECS 2017 is a competency-based model of English ability articulated through four interrelated domains, which are presented as mutually reinforcing components of a holistic construct rather than as separable skills (MoE, 2020, p. 4). Specifically, the framework comprises the following dimensions: (1) *Language Competence* (LC), referring to students' ability to understand and express meaning through listening, speaking, reading, viewing, and writing in social contexts, as well as the development of language awareness as the foundation of English learning; (2) *Cultural Awareness* (CA), concerning students' understanding of both Chinese and foreign cultures and their cognitive, attitudinal, and behavioral dispositions toward intercultural communication; (3) *Thinking Quality* (TQ), referring to students' capacity for logical, critical, and creative thinking as manifested through language use, supporting problem-solving and judgment across contexts; and (4) *Learning Ability* (LA), relating to students' awareness and use of learning strategies, their ability to access learning resources through multiple channels, and their capacity for autonomous and lifelong English learning.

The ECS 2017 presents these four domains as a curriculum policy framework, not as a detailed account of language ability theory. In explaining the revision process, the document states that the revision involved scientific review, research, testing, consultation, and expert input (MoE, 2020, p. 3). It also notes that the English curriculum content and teaching approach were informed by international comparison and a review of disciplinary theories (MoE, 2020, p. 8). In this article, *LC*, *CA*, *TQ*, and *LA* are used as policy constructs for examining NMET alignment, rather than as a separate set of theoretical constructs to be validated independently.

In addition to the four competency domains in the ECS 2017, this study includes Construct Representation (CR) as an additional analytical category. *CR* is not treated as a fifth ECS competency domain. The category draws on the ECS 2017 guidance on assessment and is used to examine whether NMET tasks represent the competencies emphasized in the ECS 2017 through features such as integrated language use, higher-order reasoning, contextual relevance, and task authenticity.

The four ECS competency domains present English proficiency as an integrated and multidimensional construct encompassing linguistic, cognitive, and socio-cultural dimensions of language use. This integrated framing establishes policy expectations that extend beyond the assessment of discrete language knowledge. At the same time, post-reform scholarship has noted that not all competencies emphasized in the curriculum are equally amenable to representation within standardized assessment formats, particularly under high-stakes conditions that prioritize reliability, comparability, and scoring efficiency (Tang et al., 2025; Wang & Luo, 2019; Zou, 2018). In this respect, the ECS 2017 provides the policy construct framework for examining alignment between curriculum intentions and assessment design.

2.3 The NMET in relation to curriculum reform

As the English component of the *Gaokao*, the NMET occupies a central position in China's secondary education system. Beyond its role as a subject test, the examination functions as an institutional mechanism through which educational opportunities are distributed, situated at the intersection of meritocratic selection and broader concerns with equity (Zou, 2018). Within the context of ongoing *Gaokao* reforms, including diversification of subject combinations and adjustments to admission policies, the NMET continues to exert a strong influence on curriculum enactment and instructional priorities in senior high schools (Zhu et al., 2026).

Since the promulgation of the ECS 2017, the NMET has increasingly been framed within policy discourse as an assessment expected to support competency-oriented curriculum reform. At the same time, existing NMET research has often examined the test through its effects on teaching and learning. This focus is evident in empirical studies of student learning experiences, instructional practices, and affective dimensions such as test anxiety (Zhang & Bournot-Trites, 2021), as well as in Rong and Yi's (2021) review of 36 NMET washback studies published between 2011 and 2020 and retrieved from China National Knowledge Infrastructure (CNKI). Recent studies further indicate that these effects are not uniform, with test takers from different provinces reporting divergent perceptions of the examination's influence on learning outcomes (Zhang, 2024). Less direct attention has been given to how NMET task design represents the constructs and competencies emphasized in curriculum policy.

The NMET itself has therefore received less direct attention as a site where curriculum-oriented competencies are represented, emphasized, or constrained through test design. Even when curriculum-test relationships are addressed, analyses have often relied on broad comparisons between stated curriculum objectives and test components, rather than systematic examination of how policy constructs are represented across test sections and task types (Wallace & Ke, 2023).

This focus on consequences and instructional impact has left issues of construct representation within the examination itself less clearly specified (Zou, 2018). Against this background, the NMET remains a consequential case for examining how competency-oriented curriculum reforms are translated into high-stakes assessment design. Despite its central role in policy discourse, the extent to which curriculum goals articulated in the ECS 2017 are reflected in the construct representation of the NMET has yet to be sufficiently specified. This question is especially salient in provinces such as Zhejiang, where the NMET operates within a dense landscape of curriculum and examination reforms, making construct representation important for understanding how national curriculum intentions enter high-stakes assessment practice.

2.4 Research gap and the present study

Existing research has documented the ambitions of China's competency-oriented English curriculum reform and the central role of the NMET in mediating policy enactment through high-stakes assessment. While studies have examined curriculum reform discourse, assessment reform initiatives, and the pedagogical and social consequences of the NMET, relatively little empirical work has systematically examined how curriculum-defined competencies are represented within the examination itself at the level of test design. This absence of design-level evidence limits the extent to which claims about policy enactment through the NMET can be empirically substantiated.

In response, the present study examines how the senior high school English education goals articulated in the ECS 2017 are reflected in the design, content, and construct representation of the NMET. To examine this question within a clearly defined context, the study adopts Zhejiang Province as the focal analytical context. As one of the earliest pilot regions for the "New Gaokao" reform (Zhu et al., 2026), Zhejiang has been at the forefront of implementing curriculum and assessment initiatives associated with competency-oriented education policies. At the same time, the NMET papers administered in Zhejiang in January and June 2025 follow the *National Unified Examination Paper Volume I* (NUEP-1, 新高考全国1卷) and adhere to national test specifications, allowing alignment to be examined within a locally implemented but nationally standardized assessment framework. This provincial focus allows this study to examine how national curriculum goals are represented in a locally implemented but nationally standardized examination. On this basis, this study addresses the following research question:

RQ. How are the senior high school English education goals articulated in ECS 2017 represented in the design, content, and construct representation of the NMET in Zhejiang Province, China?

3 Methodology

3.1 Research design and analytical framework

This study adopts a validity-oriented documentary analysis design to examine how national English education policy goals are represented in the design, content, and construct of the NMET in Zhejiang

Province. Rather than treating policy-test alignment as a matter of surface correspondence between curriculum content and test components, the analysis approaches alignment as an interpretive question concerning what claims about policy-defined competencies can reasonably be supported by evidence from test design and task content.

The analytical framework is informed by Kane's (2006, 2013) argument-based approach to validation and by the Interpretation/Use Argument (IUA). Following Chapelle et al. (2008), this study treats evidence concerning the relationship between test tasks and the intended construct as construct-focused evidence, while reserving extrapolation for claims about performance beyond the test. The analysis does not use external performance data to test whether NMET scores predict students' performance beyond the examination. Instead, it examines whether the test design, content, and task structure provide a representational basis for interpreting the competencies emphasized in the ECS 2017. In this limited sense, the findings may inform the plausibility and limits of extrapolation claims, but they do not by themselves validate such claims. Alignment is therefore examined primarily at the level of construct representation, with attention to how the competencies emphasized in the ECS 2017 are represented as section-level and task-level demands across the NMET papers.

3.2 Documentary data sources

This study draws on two primary sources of documentary data: the ECS 2017 (MoE, 2020) and the NMET papers administered in Zhejiang Province in 2025.

The ECS 2017 serves as the policy data source for identifying the curriculum competencies and assessment guidance used in the analysis. Relevant statements were extracted from sections on curriculum goals, content requirements, academic quality standards, and examination guidance. These statements were grouped under the four ECS competency domains, while assessment-related statements were used to inform the additional CR category.

The January and June 2025 NMET papers administered in Zhejiang Province constitute the assessment data for this study. Zhejiang normally administers the NMET twice within an academic year, with the higher score counting toward the final *Gaokao* total. Both papers follow the national test structure associated with the ECS 2017 and comprise the same four parts: listening, reading, language use, and writing. Given their identical section order, task types, item or task numbers, and score allocation, they are treated as parallel versions of the same assessment design. Table 1 summarizes their shared structure.

Table 1
Shared Structure of the 2025 NMET Papers Administered in Zhejiang Province

Part	Task type	Items	Score
Listening (L)	Section A: Short conversations	5	7.5
	Section B: Long dialogues and monologues	15	22.5
Reading Comprehension (R)	Section A: Four reading passages with MCQs	15	37.5
	Section B: Seven-option, five-gap matching task	5	12.5
Language Use (U)	Section A: Lexical cloze	15	15
	Section B: Grammar fill-in	10	15
Writing (W)	Section A: Formal email or letter	1	15
	Section B: Narrative continuation writing	1	25
Total		67	150

Note. The abbreviations L, R, U, and W are used in the coding and expert rating materials.

3.3 Analytical procedure

The analysis proceeded in three main stages: coding the ECS 2017, coding the NMET papers, and comparing the two sets of evidence.

The analysis began with researcher-led coding of the ECS 2017 to identify policy-defined competencies as analytic units. The clause list was selective rather than exhaustive, and relevant statements were extracted from sections on curriculum goals, content requirements, academic quality standards, and examination guidance. Drawing on the four officially specified competency domains, namely *LC*, *CA*, *TQ*, and *LA*, these statements were segmented into 25 clause-level units and grouped under the four competency domains, while assessment-related statements were used to inform the *CR* category. The coding scheme and rating format are provided in Appendix 1. The clause list and domain coding were checked against the ECS 2017 domain definitions and included in the expert judgment materials used for the alignment review.

The NMET papers administered in Zhejiang Province in January and June 2025 were then coded at the level of test sections and task types. Rather than analyzing individual items in isolation, tasks were examined in terms of their dominant task demands within each section, including listening, reading comprehension, language use, and writing. For traceability, ECS clauses were labelled by domain and clause number, for example ECS-LC-01, while NMET units were labelled by year, month, section, and item or task number, for example NMET-2025-Jan-R-A-01.

The ECS-derived competency and assessment statements were then compared with patterns of task demands identified in the NMET papers. Analysis was conducted at the section and task-type levels, with attention to the consistency of competency representation across the two test administrations. *CR* was therefore used as an analytical category for examining construct representation, not as direct evidence that NMET performance extrapolates beyond the examination.

3.4 Analytical rigor and transparency

3.4.1 Expert Judgment and Agreement

Five independent experts participated in the expert judgment process. The panel included a senior high school English teacher from Zhejiang Province and four experts with backgrounds in ECS 2017 interpretation, large-scale English testing, communicative task representation, and educational measurement. Table 2 summarizes the professional backgrounds of the expert panel.

Table 2
Expert Panel for the Alignment Review

Expert	Professional background	Relevance to this study
Expert 1	Associate professor at a teacher education university	ECS 2017 interpretation and curriculum policy enactment
Expert 2	Lecturer in applied linguistics	High-stakes English testing and test validation
Expert 3	Senior high school English teacher in Zhejiang Province (>15 years)	Classroom pedagogy, curriculum enactment, and NMET preparation
Expert 4	University lecturer in applied linguistics in eastern China	Language use, discourse demands, and communicative task representation
Expert 5	Researcher in educational measurement and evaluation at a provincial research institute	Alignment indices, assessment design, and inter-rater agreement

Before rating, all experts received the same rating package, including the ECS clause list, NMET task descriptions, rating instructions, and a standardized rating sheet. The rating sheet served as the expert judgment instrument and used five alignment categories: *Strong*, *Moderate*, *Partial*, *Limited*, and *Absent*. Operational descriptions for these categories were provided in the rating sheet in Appendix 1. Experts were asked to consider the targeted competency, task demand, language activity, cognitive demand, communicative context, and observable assessment evidence when assigning ratings. No consensus-based norming session was conducted, because this study aimed to preserve independent expert judgments. Standardization was instead supported through common written instructions and rating criteria.

The initial NMET task coding was conducted by the researcher. The NMET task descriptions and preliminary alignment results were included in the expert rating package so that experts could review the relationship between NMET task demands and ECS competency units. Experts rated whether each ECS competency unit was represented in the relevant NMET section or task type. The expert task was therefore an evaluation of researcher-prepared ECS–NMET mappings, rather than a separate item-by-item coding of all 67 test items or tasks.

Inter-rater agreement was examined using Fleiss' κ (Fleiss, 1971) to indicate the stability of interpretive judgments across raters ($\kappa > .75$ = excellent agreement, .40–.75 = fair to good agreement, $\kappa < .40$ = poor agreement; Fleiss et al., 2003). Divergent evaluations were retained as rating distributions rather than revised through consensus. These distributions were then summarized for the alignment indices using the pre-specified decision rule.

3.4.2 Adapted Alignment Indices

The expert ratings were summarized through adapted Webb alignment indices, including categorical concurrence (CC), range-of-knowledge correspondence (RKC), and balance of representation (BOR) (Webb, 1997, 1999). These indices were used to describe three aspects of construct representation in the NMET: presence, breadth, and proportional distribution. They were treated as structured summaries of alignment patterns, not as definitive measures of alignment quality. In relation to the IUA, these indices were used to summarize construct representation evidence that informs how NMET task performance can be interpreted in relation to ECS-defined constructs, rather than serving as stand-alone evidence of test validity.

Before the indices were calculated, the five expert rating categories were consolidated into three alignment classifications. "Strong" and "Moderate" ratings were treated as "Aligned," "Partial" ratings as "Partially aligned," and "Limited" and "Absent" ratings as "Not aligned." Divergent expert ratings were retained as rating distributions and then summarized using this pre-specified decision rule. When ratings were evenly split, the more conservative classification was used. The resulting consolidated alignment matrix served as the input for the adapted Webb alignment indices.

CC was used to examine whether each ECS domain was represented in the NMET. A domain was counted as present when at least one clause in that domain was classified as Aligned or Partially aligned. RKC was used to examine coverage breadth within each domain and was calculated as the ratio of addressed clauses to the total number of clauses in that domain:

$$\mathrm{RKC} = \frac{O_{\mathrm{addressed}}}{O_{\mathrm{total}}}$$

where $O_{\mathrm{addressed}}$ refers to clauses classified as "Aligned" or "Partially aligned," and O_{total} refers to the total number of clauses in the domain.

BOR was used to examine whether represented clauses within a domain were distributed proportionally rather than concentrated in a small number of test sections. Because NMET sections carry different score weights, this study used a weighted BOR index adapted from Webb's framework and following the modification proposed by Flowers et al. (2006) for assessments with unequal item or task values. The weighted index was used because representation in a higher-weighted section contributes more to the overall assessment profile than representation in a lower-weighted section. For example, a domain represented across listening, reading, language use, and writing would receive a higher value than a domain represented mainly in reading and one writing task, because the former shows a more even distribution across the test structure. The formula used was:

$$BOR_w = 1 - \frac{\sum_{k=1}^{O}\left|\frac{S_k}{T} - \frac{1}{O}\right|}{2\left(1\text{-}\frac{1}{O}\right)}$$

where O denotes the number of represented clauses within a domain, S_k represents the weighted score associated with the k-th represented clause, and T is the total weighted score across all represented clauses in that domain. BOR_w values around or above .70 were interpreted as indicating acceptable proportionality (Webb, 1999), while lower values were interpreted as indicating uneven distribution.

4 Findings

4.1 Preliminary domain-level mapping

Table 3 presents the preliminary mapping used to organize the expert rating process. The table identifies NMET sections and task types selected for expert review because of their possible relevance to the ECS 2017 domains. It does not report final alignment judgments. At this stage, the mapping served only to identify possible relationships between ECS domains and NMET task features for expert review. The strength and extent of these relationships are examined in the subsequent analyses.

Table 3
Representation of ECS 2017 Competency Domains in the 2025 NMET

ECS Domain	NMET Sections/Task Types Identified for Review	Preliminary Mapping Rationale
Language Competence (LC)	Listening, Reading, Language Use, Writing	Sections involve listening, reading, language use, or written expression
Cultural Awareness (CA)	Selected Reading passages; conditional Writing prompts	Texts or prompts were included when they involved cultural content or intercultural interpretation
Thinking Quality (TQ)	Reading inference and analysis items; Continuation writing	Tasks were included when they involved inference, reasoning, analysis, or narrative development
Learning Ability (LA)	None	No direct task evidence was found for learning strategy use, self-regulation, or reflective learning
Construct Representation (CR)	Listening and Writing sections; selected Reading tasks	Sections were included when they involved contextualization, discourse processing, authenticity, or integrated language use

4.2 Expert-validated alignment evidence

4.2.1 Expert rating distributions by domain

This section reports the expert rating distributions for the alignment matrix. As described in the methodology section, five experts rated whether each ECS competency unit was represented in the relevant NMET section or task type, using the five-category alignment scale. The 25 ECS competency units were distributed across the five analytical domains as follows: LC = 6 clauses, CA = 4 clauses, TQ = 4 clauses, LA = 4 clauses, and CR = 7 clauses, producing 125 expert ratings in total. Table 4 reports the aggregated distribution of expert judgments across the five analytical domains.

Table 4
Distribution of Expert Judgments Across Rating Categories (N = 5 Experts)

ECS Domain	Strong	Moderate	Partial	Limited	Absent	Total
LC	17	13	0	0	0	30
CA	0	1	10	5	4	20
TQ	0	3	9	5	3	20
LA	0	0	1	5	14	20
CR	6	20	9	0	0	35

A clear domain gradient emerged. For LC, all 30 judgments (6 clauses × 5 experts) fell within the top two categories, comprising 17 "Strong" and 13 "Moderate" ratings, with no "Partial," "Limited," or "Absent" judgments. This concentration indicates stable expert endorsement that the NMET systematically represents LC-related constructs within the tested modalities of listening, reading, language use, and writing. For example, LC-03, which concerns the recognition and construction of coherent discourse structures, was mapped to Reading Section B, items 36–40, a seven-option, five-gap sentence insertion task. Experts rated this alignment as 3 "Strong" and 2 "Moderate," indicating broad agreement that the task directly assessed discourse coherence.

By contrast, CA judgments were dispersed and largely below the upper categories. Of 20 ratings, "Partial" was most frequent (10), followed by "Limited" (5) and "Absent" (4), with only one "Moderate" rating and no "Strong" ratings. This distribution suggests that cultural themes are recognized in the test, but their representation is selective and not structurally embedded. For instance, CA-01, which concerns knowledge of Chinese and global cultures, was mapped to selected reading texts with social or cultural themes. Expert ratings were concentrated in the lower middle range, with 4 "Partial" and 1 "Limited" ratings. This pattern suggests that some texts provided cultural or social content, but the associated tasks mainly required comprehension, inference, or identification of textual information rather than direct demonstration of cultural awareness.

A similar pattern was observed for TQ, though with slightly more favorable evaluations than CA. Among 20 judgments, experts assigned 3 "Moderate" and 9 "Partial" ratings, alongside 5 "Limited" and 3 "Absent" ratings. This profile indicates that inferential or analytical demands are identifiable in some tasks, but higher-order dimensions are unevenly represented.

For LA, the distribution clustered at the lowest categories. Of 20 judgments, 14 were "Absent" and 5 were "Limited," with only one "Partial" rating and none in the top two categories. This pattern indicates broad expert convergence that LA-related constructs lie largely outside the scope of NMET task design. A clear example is LA-02, which concerns students' ability to plan and regulate their own English learning. This clause received 5 "Absent" ratings because no NMET task required students to demonstrate planning, monitoring, self-regulation, or reflective adjustment.

Finally, CR showed a comparatively positive but differentiated profile. Across 35 judgments, experts assigned 6 "Strong" and 20 "Moderate" ratings, alongside 9 "Partial" ratings, with no "Limited" or

"Absent" judgments. This suggests that several construct representation features emphasized in ECS 2017, such as contextualized language use and integrated task demands, were perceived as reasonably reflected in the NMET, while some CR descriptors remained only partially realized across sections.

4.2.2 Expert agreement

Inter-rater agreement across the five experts was examined using Fleiss' kappa (κ), based on the complete set of 25 ECS clauses and 125 clause-level ratings. The recalculated coefficient was κ = .366, with an average observed agreement of .504 and an expected agreement of .218. The calculation procedure is detailed in Appendix 2. Using the interpretive criteria reported by Fleiss et al. (2003), this value falls below the .40 threshold for fair-to-good agreement. The κ value was therefore interpreted together with the category-level rating distribution and clause-level disagreement patterns. Across the 125 ratings, the category distribution was Strong = 18.4%, Moderate = 29.6%, Partial = 23.2%, Limited = 12.0%, and Absent = 16.8%.

Because κ summarizes agreement across all domains and rating categories, it does not show where disagreements occurred or whether they involved adjacent or distant rating levels. The clause-level examples reported above therefore help clarify the nature of the disagreement patterns. LC-03 showed disagreement only within the upper categories, with 3 "Strong" and 2 "Moderate" ratings. CA-01 showed disagreement between adjacent lower categories, with 4 "Partial" and 1 "Limited" ratings. LA-02 showed full agreement on absence, with 5 "Absent" ratings. These patterns indicate that the lower κ value reflected differences in the strength of interpretation across constructs, rather than uniform disagreement across all domains.

For the adapted Webb alignment indices, the five-category ratings were then consolidated into three classifications: "Aligned," "Partially aligned," and "Not aligned." Divergent ratings were retained as distributions rather than resolved through consensus. This consolidated matrix served as the input for the Webb index results reported next.

4.3 Webb's alignment indices

This section presents the domain-level results from the adapted alignment indices based on the expert-rated alignment matrix. The results show whether each analytical domain was represented in the NMET, how broadly each domain was covered, and how proportionally represented content was distributed across test sections.

4.3.1 Categorical concurrence (CC)

Categorical concurrence summarized whether each analytical domain had at least one clause with observable representation in the NMET.

Table 5
Categorical Concurrence Across ECS Domains

ECS Domain	Aligned Clauses	Partially Aligned Clauses	Not Aligned Clauses	Concurrence Achieved?
LC	6	0	0	Yes
CA	0	2	2	Yes
TQ	0	2	2	Yes
LA	0	0	4	No
CR	3	4	0	Yes

As Table 5 shows, LC, CA, TQ, and CR met the concurrence criterion, whereas LA did not. LC showed the clearest presence, with all six clauses classified as aligned. CR was also represented, with all seven clauses classified as either aligned or partially aligned. CA and TQ met the criterion only through partially aligned clauses, indicating selective rather than consistent representation. LA showed no categorical concurrence because none of its four clauses was linked to observable NMET task demands.

4.3.2 Range-of-knowledge correspondence (RKC)

The RKC results show differences in the breadth of domain coverage. As Table 6 indicates, LC and CR reached full correspondence, with all clauses addressed by at least one NMET section or task type. CA and TQ showed partial correspondence, with two of four clauses addressed in each domain. LA showed no correspondence, as none of its four clauses was represented in the NMET papers. These results suggest that LC and CR were broadly covered, CA and TQ were selectively covered, and LA was not represented in observable task demands.

Table 6
Range-of-Knowledge Correspondence by ECS Domain

ECS Domain	Total Clauses	Aligned	Partially Aligned	Clauses Addressed	RKC Value
LC	6	6	0	6	1.00
CA	4	0	2	2	0.50
TQ	4	0	2	2	0.50
LA	4	0	0	0	0.00
CR	7	3	4	7	1.00

4.3.3 Balance-of-Representation (BOR)

The weighted BOR results show how evenly represented competency-related content was distributed across NMET sections.

Table 7
Summary of Weighted Balance-of-Representation (BOR_w) across ECS 2017 Domains

ECS Domain	No. of Clauses (*O*)	Aligned NMET Sections	Total Weighted Score (*T*)	Value
LC	6	L, R, U, W	150.00	.77
CA	4	R-A, R-B, W-A	65.00	.56
TQ	4	R-A, W-B	62.50	.60
LA	4	None	0.00	N/A
CR	7	L, R, U, W	N/A	N/A

Note. “L” refers to listening section, “R” to reading comprehension, “U” to language use, and “W” to writing. The “R” and “W” sections are further divided into “-A” and “-B”.

As Table 7 shows, LC had the most balanced score-weighted representation, reaching a BOR_w value of .77. This indicates that LC-related content was distributed across listening, reading, language use, and writing rather than concentrated in one section. CA and TQ were less evenly distributed. CA was represented mainly through reading and practical writing tasks, while TQ was concentrated in reading and continuation writing. LA was not assigned a BOR_w value because no clauses in this domain were represented. CR was not assigned a score-weighted BOR_w value because CR descriptors refer to section-level construct representation features rather than discrete score-bearing competency clauses.

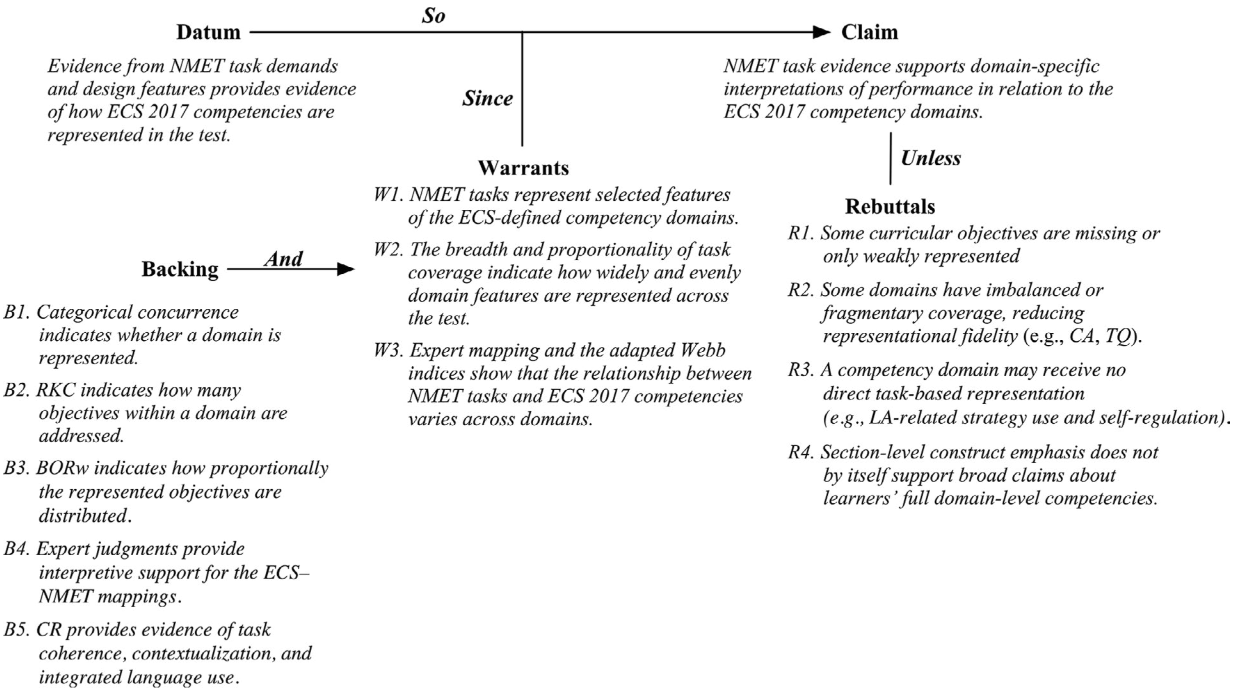
4.4 Alignment evidence and construct representation

This section synthesizes the alignment evidence reported above to clarify how NMET task demands can be interpreted in relation to the curriculum-defined competencies articulated in ECS 2017. Within the IUA, the alignment evidence is treated as construct representation evidence rather than as direct evidence of extrapolation or test validity. The expert-rated ECS-NMET mappings and adapted alignment indices jointly show the extent to which each competency domain was aligned with observable NMET task demands.

As summarized in Figure 1, the evidence was domain-specific rather than uniform. For LC, categorical concurrence, full coverage breadth, and balanced proportionality indicate strong representation across the test structure. For CA and TQ, the evidence was more limited because these domains were represented selectively and unevenly across tasks. LA lacked observable representational evidence, as no clauses in this domain were linked to NMET task demands. CR provided construct representation evidence related to task coherence, contextualization, and integrated language use, but it did not compensate for domain-level absences or uneven representation.

Figure 1

Construct Representation Evidence Map Based on Alignment Evidence



Note. Developed for this study to synthesize the alignment evidence within the present IUA analysis.

Overall, the alignment evidence suggests differentiated construct representation in the NMET. The senior high school English education goals articulated in ECS 2017 were reflected most clearly in LC and selected construct representation features, while CA and TQ were represented more selectively and LA was not directly represented in observable task demands. These findings address the research question by showing that ECS 2017 goals are reflected in the NMET in a conditional and domain-specific manner.

5 Discussion

5.1 Policy–test alignment as selective construct representation

The relationship between ECS 2017 and the NMET is not simply a matter of whether curriculum policy is reflected in the test. The findings point instead to selective construct representation. Within the NMET, some curriculum-defined competencies are made visible through observable task demands, while others are represented only indirectly or remain outside the scope of what the examination can elicit.

This selectivity is partly produced by assessment design. Curriculum standards can define broad educational goals, while high-stakes examinations translate those goals into tasks that can be administered, scored, and compared under standardized conditions. Recent work on ECS-aligned textbook implementation similarly suggests that competency-oriented policy goals are not transferred into practice in a linear manner, but are mediated by situated decisions, student needs, and contextual constraints. In the case of the NMET, this mediation occurs through assessment design. Competencies expressed through textual comprehension, discourse processing, grammatical control, and written production are therefore more readily represented in the test. Competencies involving cultural interpretation, flexible reasoning, learning strategy use, or self-regulation are harder to capture through conventional paper-based testing formats.

The uneven alignment identified in this study should therefore not be read only as a technical inconsistency between ECS 2017 and the NMET. It reflects a process in which some aspects of the curriculum are converted into assessable evidence, while others become less visible in the examination. This helps explain why LC was strongly represented, CA and TQ were selectively represented, and LA was not directly represented in observable task demands. Policy–test alignment is therefore domain-specific and conditional, shaped by the kinds of evidence that the assessment can produce.

5.2 Why construct representation differed across domains

The alignment patterns across ECS 2017 domains can be explained by the varying extent to which curriculum-defined competencies can be translated into observable, standardized, and scoreable task evidence. In a high-stakes examination such as the NMET, curriculum goals are usually translated into task demands that can be administered under common conditions and evaluated with sufficient consistency. This interpretation is consistent with validity arguments that require score-based interpretations to be supported by evidence from assessment tasks (Messick, 1989; Kane, 2013).

Language competence (LC) fits more easily within established language testing formats. Listening, reading comprehension, language use, and writing all generate observable responses within the existing NMET structure. This helps explain why LC was more systematically represented: linguistic comprehension, discourse processing, grammatical control, and written production are easier to convert into standardized test tasks than competencies requiring situated interaction or reflection.

Cultural awareness (CA) and thinking quality (TQ) present a different problem. They were not absent from the NMET, but their representation depended on selected texts, item demands, and writing prompts. CA often appeared through reading themes or communicative situations, while TQ was mainly elicited through inference, reasoning, discourse organization, or continuation writing. In such cases, the scored response remained tied to reading comprehension, textual inference, or written production, which limits how far CA and TQ can be interpreted as independently represented constructs (Chalhoub-Deville, 2016; Chapelle, 2021). Studies of ECS-aligned textbook implementation similarly show that competency-oriented goals can be expanded in classroom practice through teacher mediation, additional discussion, multimodal input, and intercultural activities (Yan & Weng, 2026). Such pedagogical expansion is less easily accommodated in a standardized examination, where broader competencies tend to be compressed into text-based and scoreable responses.

Learning ability (LA) exposes the clearest boundary of test-based construct representation. LA concerns how students plan, monitor, regulate, and reflect on their learning, but these processes are not directly visible in final test responses. Students may use learning strategies while preparing for or completing the NMET, but the examination does not require them to demonstrate those strategies as assessable evidence. This explains why LA was not represented in observable NMET task demands and is consistent with concerns that outcome-focused assessment may leave important learning processes outside the scope of direct evidence (Shepard, 1997).

These patterns suggest that differentiated construct representation is shaped by the interaction between curriculum ambition and assessment feasibility. The NMET does not simply mirror ECS 2017. It filters the curriculum through the practical requirements of high-stakes assessment, making competencies that can be elicited and scored more visible than those requiring extended interaction, self-regulation, or situated interpretation.

5.3 Interpretive boundaries of NMET performance evidence

The domain-specific alignment patterns indicate that NMET performance evidence should be interpreted within clear construct boundaries. NMET performance should not be read as evidence of the full range of ECS 2017 competencies. It can support interpretation only where the relevant competencies are made visible through task demands. This view is consistent with validity arguments that link score interpretation to the evidence generated by assessment tasks (Messick, 1989; Kane, 2013).

Where competencies are systematically represented through recurring task types and stable design features, NMET performance can be interpreted with greater confidence. This applies most clearly to LC, where listening, reading, language use, and writing tasks provide repeated evidence of comprehension, discourse processing, grammatical control, and written expression. For CA and TQ, the interpretive claim needs to be narrower because the evidence appeared mainly through selected reading themes, inference demands, and writing tasks. These task features support limited claims about CA and TQ, but not broader claims about the full curricular scope of these domains (Chalhoub-Deville, 2016; Chapelle, 2021).

LA marks the clearest boundary. Since no LA clauses were linked to observable NMET task demands, NMET performance should not be interpreted as direct evidence of students' planning, monitoring, self-regulation, or reflective learning. This does not mean that LA is unimportant to English learning. It means that LA falls outside the direct evidential scope of the NMET. These boundaries help prevent NMET results from being interpreted more broadly than the assessment evidence allows (Shepard, 1997).

5.4 Assessment design and the narrowing of curriculum representation

The alignment pattern also points to a narrowing effect in assessment design. ECS 2017 defines English learning through broad competency domains, but the NMET represents these domains through tasks that fit standardized administration and scoring. In this process, curriculum goals are not simply transmitted into assessment form. They are reshaped by what the test can make visible, measurable, and consequential.

This narrowing matters because the NMET is not only a measurement instrument. In a high-stakes assessment system, the competencies most visible in the examination are likely to become more visible in teaching, preparation, and institutional attention. Competencies weakly represented or absent in the test may remain important in policy, but they are less likely to carry the same practical weight in examination-oriented settings. Studies of ECS-aligned textbook implementation show a related pattern: competency-oriented goals often require teacher mediation, supplementary activities, and classroom-level expansion to become meaningful in practice (Yan & Weng, 2026). The NMET operates differently from classroom teaching, but it occupies a powerful position in the same policy ecology.

The construct representation (CR) results add an important qualification to this interpretation. They indicate that the NMET does include construct-relevant task features, such as discourse-based language use, contextualized task demands, and integrated language performance. However, these features do not remove the uneven representation identified across the competency domains. The issue, then, is not simple misalignment. The NMET provides a partial representation of ECS 2017, shaped by the demands of high-stakes assessment and by the types of evidence that can be produced within the test. Strong representation of LC, selective representation of CA and TQ, and the absence of LA suggest that the examination may reinforce a narrower public understanding of English competence than the one articulated in the curriculum standards.

5.5 Implications

The uneven alignment patterns identified in this study are relevant to assessment developers, curriculum policymakers, teachers, schools, and researchers concerned with policy-test alignment.

For assessment developers and examination authorities, the main implication concerns the interpretation of NMET evidence. The NMET provides stronger evidence for LC because this domain is repeatedly elicited through observable task demands. Claims about CA, TQ, and LA, however, need to remain more qualified because these domains were selectively represented or not directly represented in the assessment. NMET performance should not be treated as comprehensive evidence of students' attainment of all competency goals in ECS 2017.

For curriculum policymakers, teachers, and schools, the findings point to the limits of relying on examination preparation as evidence of curriculum implementation. The partial representation of CA and TQ, together with the absence of LA, suggests that broad competency goals can be narrowed when translated into high-stakes assessment tasks. These competencies remain important, but they require pedagogical attention beyond the task demands most visible in the NMET. Closer coordination is therefore needed between policy intentions, assessment design, classroom teaching, and the evidential claims attached to test performance.

Methodologically, this study offers a validation-oriented way of examining policy-test alignment. By combining expert judgment, adapted alignment indices, and an argument-based validity perspective, the analysis identifies whether each ECS domain was classified as "Aligned," "Partially aligned," or "Not aligned" with observable NMET task demands. It also clarifies the evidential boundaries within which assessment-based interpretations can be justified.

5.6 Limitations

The absence of a speaking component in the Zhejiang NMET papers analyzed in this study limits the interpretation of LC. Constructs related to oral production and spoken interaction could not be examined through the present alignment framework, despite their inclusion in ECS 2017. The strong representation of LC reported in this study should therefore be understood as referring to the tested modalities of listening, reading, language use, and writing, rather than to the full range of communicative language abilities described in the curriculum standards.

This study is also limited to the January and June 2025 NMET English papers used in Zhejiang Province. This focus is analytically meaningful given Zhejiang's role in the "New Gaokao" reform and the use of the *National Unified Examination Paper Volume I* (NUEP-1). Even so, the findings should not be read as evidence of stable alignment patterns across years, future administrations, or provinces with different assessment arrangements. Further work could extend the analysis across a longer period or compare provinces using the same or different examination papers.

Because alignment judgments involve construct interpretation, the expert-based component of the analysis carries some interpretive uncertainty, especially where evidence was indirect. Alignment judgments cannot be produced mechanically from test content, but expert ratings may still vary with professional background, familiarity with the task type, interpretation of the construct, or rating severity. This concern is consistent with language assessment research showing that expert ratings can be affected by rater background, rating severity, and differences in construct interpretation (Eckes, 2012). In this study, the most uncertain cases involved CA in culturally themed reading texts, TQ in inference or continuation-writing tasks, and CR in construct-relevant task features. These cases required experts to decide whether a task contained relevant content only at the thematic level or directly elicited evidence of the target construct. Future studies could examine such judgment processes more closely through larger expert panels, post-rating interviews, rater training, or rater-effect analysis.

6 Conclusion

This study examined policy-test alignment in the NMET as a question of construct representation, focusing on whether task-based assessment evidence can be interpreted in relation to curriculum-defined competencies articulated in ECS 2017. Rather than treating alignment as a matter of surface correspondence, the analysis approached it as a validity problem concerning the evidential conditions under which NMET performance can support curriculum-related interpretations.

The findings suggest that alignment between curriculum intentions and assessment design is not uniform across competency domains. Language competence (LC) was systematically instantiated through recurring task types and stable design features, providing stronger evidence for interpretation within this domain. By contrast, cultural awareness (CA) and thinking quality (TQ) were represented selectively and unevenly, which constrains the scope of defensible interpretation. Learning Ability (LA) was not instantiated in observable task demands and therefore falls outside the direct evidential scope of the NMET papers analyzed in this study. These findings show that the senior high school English education goals articulated in ECS 2017 are represented in the NMET through selective task-based evidence rather than comprehensive coverage of all competency domains.

These patterns clarify what NMET performance can and cannot reasonably be interpreted to indicate in relation to curriculum-defined competencies. Alignment evidence does not justify broad claims about curriculum enactment or policy realization. Instead, it delineates the evidential boundaries within which assessment-based interpretations remain valid. In this sense, this study highlights the importance of examining what assessment evidence can and cannot support in high-stakes testing contexts. Policy-test alignment should therefore be treated as a validity issue, not simply as a comparison between curriculum content and test content.

Methodologically, this study demonstrates a practical approach to alignment analysis by combining the IUA, adapted Webb alignment indices, and expert judgment within a single framework. This combination made it possible to examine not only whether ECS 2017 competencies appeared in the NMET, but also how strongly, broadly, and evenly they were represented across test sections and task types. In practice, NMET evidence should be read in domain-specific terms. The NMET provides stronger evidence for some competencies than for others, and its scores should therefore be interpreted carefully rather than treated as evidence of the full curriculum framework. For competencies that are weakly represented or absent in the NMET, classroom-based or formative evidence remains necessary. Future research could extend the analysis across multiple years, provinces, and assessment formats, particularly where speaking assessment is incorporated, and could further examine how expert judgment shapes alignment interpretations.

Declaration of Generative AI Usage

In the preparation of this study, Grammarly was used to assist with grammar checking and minor language refinement. After using this tool, the authors reviewed and edited the manuscript as needed and take full responsibility for the content of the publication.

Appendix

Appendix 1. Expert Rating Sheet

ECS Clause ID	Abbreviated Description	NMET Section(s) Mapped	Strong	Moderate	Partial	Limited	Absent
LC-01	Express meaning in spoken/written discourse	L-A/B; R-A/B; W-A	☐	☐	☐	☐	☐
LC-02	Comprehend texts of moderate complexity	R-A-21-35	☐	☐	☐	☐	☐
LC-03	Discourse coherence recognition	R-B-36-40	☐	☐	☐	☐	☐
LC-04	Lexical/grammatical accuracy	U-A/B	☐	☐	☐	☐	☐
LC-05	Practical writing	W-A-66	☐	☐	☐	☐	☐
LC-06	Extended written expression	W-B-67	☐	☐	☐	☐	☐
CA-01	Cultural knowledge awareness	R-B passage	☐	☐	☐	☐	☐
CA-02	Intercultural communication	R-A selected	☐	☐	☐	☐	☐
CA-03	Cultural confidence	W-A	☐	☐	☐	☐	☐
CA-04	Cultural dissemination	-	☐	☐	☐	☐	☐
TQ-01	Logical inference	R-A inference	☐	☐	☐	☐	☐
TQ-02	Analytical thinking	R-A analysis	☐	☐	☐	☐	☐
TQ-03	Critical evaluation	-	☐	☐	☐	☐	☐
TQ-04	Creativity/problem solving	W-B-67	☐	☐	☐	☐	☐
LA-01	Strategy use	-	☐	☐	☐	☐	☐
LA-02	Self-regulated learning	-	☐	☐	☐	☐	☐
LA-03	Reflective learning	-	☐	☐	☐	☐	☐
LA-04	Lifelong learning orientation	-	☐	☐	☐	☐	☐
CR-01	Real-life communicative use	L/W tasks	☐	☐	☐	☐	☐
CR-02	Integrated skills	L + R + W	☐	☐	☐	☐	☐

CR-03	Authenticity of tasks	W-A; selected R	☐	☐	☐	☐	☐
CR-04	Discourse processing	L-A/B; R-B	☐	☐	☐	☐	☐
CR-05	Communicative performance	W-A/B	☐	☐	☐	☐	☐
CR-06	Task completion/ problem solving	W-B	☐	☐	☐	☐	☐
CR-07	Balanced difficulty, fairness	Whole test	☐	☐	☐	☐	☐

Strong = explicit and comprehensive alignment
Moderate = clear but incomplete alignment
Partial = alignment with selected aspects of the ECS clause
Limited = weak or marginal alignment with little observable task evidence
Absent = no observable alignment

Appendix 2. Calculation process

Fleiss' Kappa Calculation

N = 25 ECS clauses (rating units), k = 5 raters per clause,
the total number of ratings across clauses is $N \times k = 25 \times 5 = 125$.

Observed agreement for clause i is calculated as: $P_i = \frac{1}{k\,(k\text{-}1)} = \sum_{j=1}^{5} n_{ij}\,(n_{ij} - 1)$

e.g.,
In Clause LC-01, Ratings: 4 "*Strong*," 1 "*Moderate*," 0 ratings in all other categories.

$P_{\text{LC-01}} = \frac{1}{5\times4}\,[4(4\text{-}1)+1(1\text{-}1)+0(0\text{-}1)+0(0\text{-}1)+0(0\text{-}1)] = \frac{3}{5} = 0.600;$

This calculation was applied to all 25 clauses, yielding an average observed agreement: $\overline{P} = \frac{1}{25}\sum_{i=1}^{25} P_i = 0.504;$

Thus, approximately 50.4% of all potential rater pairs agreed on the same category on average.
The overall category proportions were calculated as follows:

Category	Proportion p_j
Strong	.184
Moderate	.296
Partial	.232
Limited	.120
Absent	.168

Expected agreement is calculated as: $\overline{P}_e = \sum_{j=1}^{5} p_j^2$

$\overline{P}_e$ = 0.218, thus 21.8% agreement would be expected by chance.

Fleiss' $\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} = \frac{0.504 - 0.218}{1 - 0.218} = 0.366$

A κ value of .366 falls below the .40 threshold for fair-to-good agreement under the interpretive criteria reported by Fleiss et al. (2003). For this reason, the coefficient was interpreted together with the category-level rating distribution and clause-level disagreement patterns, rather than treated as sufficient evidence of strong agreement.

Weighted Balance-of-Representation (BOR_w)

Because NMET sections differ greatly in their scoring weights, the present study adopts a weighted version of the index, following Flowers et al. (2006). Item counts are replaced with score-based allocations, and expert-assigned correspondence strengths are incorporated. The weighted index is defined as:

$$BOR_w = 1 - \frac{\sum_{k=1}^{O} \left| \frac{S_k}{T} - \frac{1}{O} \right|}{2\left(1 - \frac{1}{O}\right)}$$

S_k is the total weighted score of NMET tasks aligned with the k-th objective within a domain, and $T = \sum_{k=1}^{O} S_k$ is the total weighted score across all aligned objectives in that domain.

LC: $BOR_w = 1 - 0.2276 = 0.7724 \approx 0.77$ **(>0.70)**
CA: $BOR_w = 1 - 0.4359 = 0.5641 \approx 0.56$ **(<0.70)**
TQ: $BOR_w = 1 - 0.40 = 0.60$ **(<0.70)**
LA: $\mathrm{BOR_w} = 1 - \frac{\sum_{k=1}^{O} \left| \frac{S_k}{T} - \frac{1}{O} \right|}{2\left(1 - \frac{1}{O}\right)}$ **(not computable)**

References

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, *33*(4), 453–472. https://doi.org/10.1177/0265532215593312

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright & J. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language*™ (pp. 319–352). Routledge. https://doi.org/10.4324/9780203937891

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright & J. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language*™ (pp. 1–25). Routledge. https://doi.org/10.4324/9780203937891

Chapelle, C. A. (Ed.). (2021). What is argument-based validity? In *Argument-based validation in testing and assessment* (pp. 1–19). SAGE Publications. https://doi.org/10.4135/9781071878811

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*, *25*(1), 15–37. https://doi.org/10.1177/0265532207083743

Cheng, Y., & Hamid, M. O. (2025). Social impact of Gaokao in China: A critical review of research. *Language Testing in Asia*, *15*(22), 1–17. https://doi.org/10.1186/s40468-025-00355-y

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Fleiss, J. L., Levin, B., & Paik, M. C. (Eds.). (2003). The measurement of interrater agreement. In *Statistical methods for rates and proportions* (3rd ed, pp. 598–626). John Wiley & Sons, Inc. https://doi.org/10.1002/0471445428

Flowers, C., Browder, D., & Ahlgrim-Delzell, L. (2006). An analysis of three states' alignment between language arts and mathematics standards and alternate assessments. *Exceptional Children*, *72*(2), 201–215. https://doi.org/10.1177/001440290607200205

Gao, Y. (2018). China's fluctuating English education policy discourses and continuing ambivalences in identity construction. In A. Curtis & R. Sussex (Eds.), *Intercultural communication in Asia: Education, language and values* (pp. 241–261). Springer Cham. https://doi.org/10.1007/978-3-319-69995-0_12

Gil, J. (2016). English language education policies in the People's Republic of China. In R. Kirkpatrick (Ed.), *English language education policy in Asia* (pp. 49–90). Springer. https://doi.org/10.1007/978-3-319-22464-0_3

Gottlieb, M., & Chapman, M. (2022). Language standards frameworks and language assessment: A codependent relationship. *International Journal of TESOL Studies*, *4*(1), 48–68. https://doi.org/10.46451/ijts.2022.01.05

Hsieh, C. N. (2024). Building a validity argument for the TOEFL Junior® Tests. *ETS Research Report Series*, *2024*(1), 1–22. https://doi.org/10.1002/ets2.12379

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed, pp. 17–64). Rowman & Littlefield Publishers.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

MoE (Ministry of Education of the People's Republic of China). (2020). *Senior High School English Curriculum Standards* (2017 edition, revised in 2020). People's Education Press.

Pan, L., & Wang, Q. (2024). Using English to realize "the Chinese dream": The political-economic ideology of language planning in China. *Asia Pacific Journal of Education*, 1–15. https://doi.org/10.1080/02188791.2024.2418819

Pan, M., & Qian, D. D. (2017). Embedding corpora into the content validation of the grammar test of the national matriculation English test (NMET) in China. *Language Assessment Quarterly*, *14*(2), 120–139. https://doi.org/10.1080/15434303.2017.1303703

Rong, L., & Yi, P. (2021). A review of studies on the washback effect of national matriculation English test (NMET) in China 2011–2020. *IRA-International Journal of Education & Multidisciplinary Studies*, 17(3), 186–196. https://dx.doi.org/10.21013/jems.v17.n3.p10

Saeedi, Z., Tajeddin, Z., & Tadayon, F. (2025). Assessment of English as a lingua franca and its principles: A research synthesis. *International Journal of TESOL Studies, 7*(3), 130-152 https://doi.org/10.58304/ijts.250711

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24. https://doi.org/10.1111/j.1745-3992.1997.tb00585.x

Tang, X., Liu, Y., & Pan, L. (2025). From global English to global Englishes: Implications for a new reading of the national English curriculum standards of China. *Cross-Cultural Education Studies*, *1*(2), 43–56. https://doi.org/10.63385/cces.v1i2.219

Wallace, M. P., & Ke, H. (2023). Examining the content alignment between language curriculum and a language test in China. *TEFLIN Journal*, *34*(1), 116–135. https://doi.org/10.15639/teflinjournal.v34i1/116-135

Wang, Q., & Luo, S. (2019). Shifting from teaching the subject to developing core competencies through the subject: The revised senior middle school English curriculum standards (2017 Edition) in China. In X. Gao (Ed.), *Second handbook of English language teaching* (pp. 109–134). Springer. https://doi.org/10.1007/978-3-030-02899-2_6

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education* (Research Monograph No. 6). Council of Chief State School Officers. https://eric.ed.gov/?id=ED414305

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Council of Chief State School Officers. https://eric.ed.gov/?id=ED440852

Xu, Y., Li, X., & Chen, J. (2025). Test review: Computer-based English listening and speaking test (CELST) of national matriculation English test (NMET) Guangdong version in China. *Language Testing*, *42*(2), 238–249. https://doi.org/10.1177/02655322241255712

Yan, L., & Weng, T. (2026). "We can only do addition": An ecological perspective on Chinese EFL high school teachers' agentic use of the new senior high school English textbooks. *Language Teaching Research*, 1–36. https://doi.org/10.1177/13621688261433883

Zhang, H. (2024). The washback of the national matriculation English test on senior high school English learning outcomes: Do test takers from different provinces think alike? *Language Testing in Asia*, *14*(1), 1–14. https://doi.org/10.1186/s40468-024-00286-0

Zhang, H., & Bournot-Trites, M. (2021). The long-term washback effects of the national matriculation English Test on college English learning in China: Tertiary student perspectives. *Studies in Educational Evaluation*, *68*, Article e100977. https://doi.org/10.1016/j.stueduc.2021.100977

Zhu, P., Li, M., & Zhu, Z. (2026). Diversification of subject combinations in the national college entrance examination and educational reforms in senior secondary schools: Findings from China's policies on college admissions. *Asia Pacific Education Review*, *27*, 413–424. https://doi.org/10.1007/s12564-024-09949-2

Zou, Y. (2018). Educational opportunity between meritocracy and equity: A review of the national college entrance examination in China since 1977. In B. Hamre, A. Morin & C. Ydesen (Eds.), *Testing and inclusive schooling: International challenges and opportunities* (pp. 34–46). Routledge. https://doi.org/10.4324/9781315204048

Lu Geng Sun is a Ph.D. candidate in the English as an International Language (EIL) Program at Chulalongkorn University. His research interests lie in language assessment, language policy and planning, and qualitative research methods.

ORCID ID: https://orcid.org/0009-0003-7583-8457

Jirada Wudthayagorn, Ph.D., is an associate professor, the Director of Chulalongkorn University Language Institute (CULI), a governing board member of RELC (SEAMEO Regional Language Centre), and the past President of the Asian Association of Language Assessment (AALA). She was a recipient of the Royal Thai Government Scholarship to pursue her Ph.D. studies in Applied Linguistics at the University of Pittsburgh, USA. Her research interests lie in language assessment and evaluation, language policy, and quantitative research methods.

ORCID ID: https://orcid.org/0000-0001-5992-1996