NEAS Test of English

Construct, Design, and Validation Framework

A CEFR-aligned, AI-mediated English proficiency assessment

Author: Richard Stevenson
March 2026

Executive Summary

The NEAS Test of English is a CEFR-aligned, AI-mediated language assessment designed to measure communicative proficiency across the four macro-skills of reading, listening, writing, and speaking. The test is grounded in contemporary models of language ability, which conceptualise proficiency as the ability to construct and convey meaning through integrated linguistic, cognitive, and discourse-level processes.

The assessment adopts a performance-based approach in which language ability is evaluated through task-based activities rather than through discrete-point testing of grammar or vocabulary. Linguistic competence is treated as an embedded resource within communicative performance, and proficiency is defined in terms of control, flexibility, and communicative effectiveness appropriate to CEFR levels.

A central design principle is that difficulty is determined by processing and communicative demands rather than by surface features such as text length or lexical frequency. In the receptive skills, this involves increasing demands on inferencing, integration, and discourse-level interpretation. In the productive skills, it involves increasing expectations for organisation, development, abstraction, and pragmatic appropriacy.

The test is designed to support multiple use cases, including placement, level certification, and benchmarking. To ensure accurate classification and robust construct coverage, receptive components sample performance across a range of CEFR levels, while productive components use common task sets that allow differentiation through performance quality rather than task allocation.

Scoring is aligned to CEFR through analytic criteria in productive skills and dichotomous item-based scoring in receptive skills. Level assignment is based on profile-based decision rules that require consistent performance across criteria and interpret errors in terms of their impact on meaning rather than formal correctness.

The test is delivered in an AI-mediated environment with standardised input and behaviourally anchored scoring criteria, ensuring comparability, reliability, and compatibility with automated scoring systems while maintaining alignment with communicative models of language ability.

Taken together, the NEAS Test of English provides a valid, reliable, and interpretable measure of language proficiency, grounded in contemporary theory and aligned with international best practice in CEFR-based assessment.

Table of Contents

1. Introduction 5

2. Theoretical Framework 6
   - 2.1. Communicative competence and the nature of language ability 6
   - 2.2. The CEFR and the action-oriented model of language use 6
   - 2.3. Validity as an interpretive argument 7
   - 2.4. Integration of theoretical perspectives 7

3. Test Construct and Design Principles 7
   - 3.1. Macro-skills rationale 8
   - 3.2. Embedded linguistic competence 8
   - 3.3. Progression, difficulty, and level differentiation 9
   - 3.4. Test use cases and implication for design 9
   - 3.5. Topic selection, schema activation, and cultural fairness 10
   - 3.6. Task design, comparability, and observability 11
   - 3.7. AI-mediated constraints and design implications 11
   - 3.8. Summary of construct and design principles 12

4. **Receptive Skills (Reading & Listening)** 12
   - 4.1. Receptive Skills Overview 12
   - 4.2. Reading-Specific Design Parameters 13
   - 4.2.1. Text design principles 13
   - 4.2.2. Lexical profiling and text calibration 14
   - 4.2.3. Text length, timing, and processing load 14
   - 4.2.4. Pilot task validation and text appropriateness 15
   - 4.3. Listening-Specific Design Parameters 15
   - 4.3.1. Real-time processing and delivery constraints 16
   - 4.3.2. Control of speech rate, delivery, and pausing 16
   - 4.3.3. Control of text length and cognitive load 16
   - 4.3.4. Number of listens as a construct variable 17
   - 4.3.5. Accent control and authenticity 17
   - 4.3.6. Administration model and method effects 17
   - 4.3.7. Role of lexical profiling in listening validation 18
   - 4.3.8. Summary of Receptive Skills Design Rationale 18

5. **Productive Skills (Speaking & Writing)** 18
   - 5.1. Productive Skills Overview 18
   - 5.2. Writing-Specific Design Parameters 19
   - 5.2.1. Task design and comparability 19
   - 5.2.2. Discourse organisation and idea development 19
   - 5.2.3. Lexical profiling and task calibration 19
   - 5.2.4. Timing, planning, and revision 20
   - 5.2.5. Pilot validation 20
   - 5.3. Speaking-Specific Design Parameters 20
   - 5.3.1. Task design and structure 20
   - 5.3.2. Discourse management and real-time processing 21
   - 5.3.3. Delivery conditions and standardisation 21
   - 5.3.4. Sequencing and cognitive load 21
   - 5.3.5. Pilot validation 22
   - 5.4. Summary of Productive Skills Design Rationale 22

6. Assessment and Scoring Framework 22
   - 6.1. Scoring construct and underlying principles 22
   - 6.2. Analytic scoring in productive skills 23

6.2.1. Rationale for analytic criteria 23
6.3. Comprehension-based scaling and error interpretation 23
6.4. Scoring in receptive skills 23
6.4.1. Rationale for dichotomous scoring 24
6.5. CEFR alignment and scale design 24
6.6. Combining criteria: decision rules and overall level assignment 24
6.6.1. Rationale for decision rules 24
6.7. Consistency across skills 25
6.8. Summary of scoring rationale 25
6.9. Rubric and Descriptor Design Principles 25
6.9.1. Conceptualising descriptors as measurement instruments 25
6.9.2. Observable behaviour and avoidance of abstract constructs 26
6.9.3. Functional interpretation of error and control 26
6.9.4. Integration and separation of criteria 26
6.9.5. Developmental progression and scale logic 27
6.9.6. Avoidance of native-speaker norms and academic bias 27
6.9.7. Native-speaker benchmarking 27
6.9.8. Academic discourse bias 27
6.9.9. Compatibility with AI-mediated scoring 27
6.9.10. Alignment with CEFR descriptors 28
6.9.11. Summary of descriptor design rationale 28

7. Scaling, CEFR Alignment, and Decision Rules 28
7.1. Conceptual basis for scaling and CEFR alignment 28
7.2. Band structure and developmental scale 28
7.3. Construct-based cut scores 29
7.4. Profile-based decision rules 29
7.5. Relationship between analytic profiles and overall level 30
7.6. Alignment with external frameworks (IELTS and related scales) 30
7.7. Validity and defensibility of the scaling model 30

8. Validity and Reliability Considerations 30
8.1. Construct validity 30
8.2. Fairness and avoidance of construct-irrelevant variance 31
8.3. Reliability and consistency of measurement 31
8.4. AI-mediated scoring considerations 32
8.5. Summary of validity and reliability framework 32

9. Conclusion 33

10. Reference list 33

1. Introduction

The NEAS Test of English has been developed as a comprehensive assessment of English language proficiency across the four macro-skills of reading, listening, writing, and speaking. The test is designed for use in educational and professional contexts where a reliable, interpretable, and CEFR-aligned measure of language ability is required. It aims to provide valid and consistent evidence of candidates' capacity to use English effectively in a range of communicative situations, from routine, everyday interactions to more complex and abstract domains.

The design of the test is grounded in contemporary models of communicative language ability, which conceptualise proficiency as the ability to construct and convey meaning through language use rather than as the possession of discrete linguistic knowledge. In this view, grammatical, lexical, discourse, and pragmatic competences are not treated as independent skills, but as integrated resources that are realised through performance in communicative tasks. Accordingly, the NEAS test adopts a skills-based approach, in which these components are assessed through candidates' performance in reading, listening, writing, and speaking, rather than through isolated testing of language forms.

The test is aligned with the Common European Framework of Reference for Languages (CEFR), both in its conceptualisation of language proficiency and in its reporting of outcomes. CEFR levels provide a widely recognised and internationally interpretable framework for describing what learners can do with language at different stages of development. The NEAS test operationalises these levels through carefully designed tasks and analytic scoring criteria, ensuring that reported outcomes correspond to meaningful differences in communicative ability.

In addition to its theoretical grounding, the test has been designed with particular attention to practical considerations of delivery, scoring, and use. It is intended for AI-mediated, unsupervised administration, and this has informed key design decisions, including the standardisation of task input, the use of structured response formats, and the development of scoring criteria based on observable features of performance. These decisions aim to balance the demands of reliability and scalability with the need to preserve construct validity and alignment with established models of language proficiency.

The document that follows sets out the theoretical, design, and operational foundations of the NEAS Test of English. It begins by outlining the theoretical framework underpinning the test, including models of communicative competence, CEFR alignment, and contemporary approaches to validity. It then describes the test construct and design principles, followed by detailed specifications for each skill area. Subsequent sections address scoring frameworks, scaling and CEFR alignment, validity and reliability considerations, and the design of rating descriptors. Together, these sections provide a coherent account of how the test has been developed and how its results should be interpreted.

The overarching aim of this document is to demonstrate that the NEAS Test of English is a principled, theoretically grounded, and operationally robust assessment of language proficiency. By making its design decisions explicit and situating them within established

research and practice, the document supports transparency, interpretability, and confidence in the use of test results across a range of contexts.

2. Theoretical Framework

2.1. Communicative competence and the nature of language ability

The NEAS Test of English is grounded in a communicative view of language ability, in which proficiency is understood as the capacity to use language effectively to achieve meaningful purposes in context. This perspective originates in the concept of communicative competence (Canale & Swain, 1980), which reconceptualised language ability as extending beyond grammatical knowledge to include sociolinguistic, discourse, and strategic dimensions. Within this framework, linguistic resources are not ends in themselves but serve the construction and negotiation of meaning in communicative use.

Subsequent models have further elaborated this position. Bachman (1990) conceptualises communicative language ability as comprising organisational competence (including grammatical and textual knowledge) and pragmatic competence, both of which are activated through strategic processes during task performance. A key implication of this model is that language ability cannot be reduced to discrete components such as grammar or vocabulary but emerges from the interaction of multiple knowledge systems during use (Khalifa & Weir, 2009).

These developments establish a clear theoretical position: language proficiency is most validly understood and assessed through performance in communicative tasks rather than through decontextualised testing of individual language components.

2.2. The CEFR and the action-oriented model of language use

The theoretical framework of the test is further informed by the Common European Framework of Reference for Languages (CEFR), which provides a comprehensive and widely adopted system for describing language proficiency. The CEFR adopts an action-oriented approach, defining language users as social agents who use language to accomplish tasks in specific contexts (Council of Europe, 2001, 2020). Proficiency is therefore described in terms of what learners can do with language, rather than what they know about it.

Within this model, language ability is organised around communicative activities—reading, listening, writing, and speaking—supported by underlying linguistic competences such as lexical, grammatical, and pragmatic knowledge. These competences are conceptualised as enabling resources rather than independent assessment targets, reinforcing the view that language ability is realised through performance in context (Council of Europe, 2001, 2020).

The CEFR also provides a developmental framework in which proficiency levels represent qualitative changes in the complexity, flexibility, and precision of language use. Lower levels are associated with the handling of simple, concrete tasks under strong contextual

support, while higher levels involve the ability to process and produce more abstract, nuanced, and discourse-rich language. This progression reflects increasing range, control, and independence in language use, and provides a principled basis for distinguishing levels of performance in language assessment (Council of Europe, 2001, 2020; Weir, 2005).

2.3. Validity as an interpretive argument

The interpretation and use of test scores in the NEAS Test of English are grounded in contemporary theories of validity, particularly the unified framework proposed by Messick (1989). Within this perspective, validity is not a property of the test itself, but of the inferences made from test scores. A valid assessment therefore requires a coherent and evidence-based argument that score interpretations are meaningful, appropriate, and aligned with the intended construct.

Messick's framework highlights the importance of construct representation, relevance, and the avoidance of construct-irrelevant variance. These principles require that test tasks elicit the intended language abilities while minimising the influence of factors unrelated to the construct being measured.

This general perspective is complemented by Weir's (2005) socio-cognitive model, which provides an operational account of how validity can be realised in practice. Weir identifies interacting dimensions of validity, including cognitive validity (engagement of appropriate processes), context validity (appropriateness of task conditions), scoring validity, and consequential validity.

Together, these frameworks emphasise that validity depends on the alignment between construct definition, task design, and scoring procedures. Assessment is therefore understood as an integrated system in which each component must support the intended interpretation of performance.

2.4. Integration of theoretical perspectives

Taken together, these perspectives establish a coherent theoretical foundation for the NEAS Test of English. The concept of communicative competence defines language ability as performance in context; the CEFR provides a structured and widely recognised model for describing proficiency and its development; and contemporary validity theory offers a principled basis for interpreting and justifying test scores.

This integrated framework ensures that the assessment is grounded in established theory while maintaining a clear focus on the relationship between language use, task performance, and score interpretation. It also provides the basis for subsequent design decisions, which are articulated in the following sections.

3. Test Construct and Design Principles

3.1. Macro-skills rationale

The NEAS Test of English is structured around the four macro-skills of reading, listening, writing, and speaking, reflecting the principle that communicative language ability is most validly observed through performance in these domains. Rather than treating proficiency as the accumulation of discrete linguistic knowledge, the test operationalises language ability as the capacity to integrate multiple resources in real-time comprehension and production.

Within this framework, grammar, vocabulary, discourse, and pragmatics are not treated as independent skills, but as underlying components realised through language use. The four macro-skills therefore provide the primary means of eliciting and observing performance, as they correspond to the principal ways in which language is used in real-world contexts. This approach is consistent with the CEFR's organisation of proficiency around receptive and productive activities and with the broader shift in language testing toward performance-based assessment.

Research in language testing has consistently shown that skills-based assessment is widely argued to provide stronger construct validity and greater interpretability than discrete-point testing of individual components. While tests of grammar or vocabulary may have diagnostic value, they are limited in their ability to capture communicative performance, as they often measure decontextualised knowledge rather than meaningful language use. For this reason, large-scale proficiency tests internationally have converged on a macro-skills-based structure, in which language ability is inferred from performance across reading, listening, writing, and speaking tasks (Bachman & Palmer, 1996; Purpura, 2004; Read, 2000; O'Sullivan, 2011).

The adoption of a four-skills framework in the NEAS test therefore represents a principled alignment with established theory, policy frameworks, and international assessment practice. It ensures that the test samples a broad range of communicative language use and supports meaningful interpretation of proficiency across contexts.

3.2. Embedded linguistic competence

A central design principle of the test is that linguistic competence—encompassing grammar, vocabulary, discourse, and pragmatics—is assessed as an embedded resource within communicative tasks rather than as a set of isolated constructs. This reflects the position that linguistic knowledge is only meaningful insofar as it contributes to successful language use in context (Bachman, 1990; Bachman & Palmer, 1996; Council of Europe, 2001, 2020).

This principle has direct implications for task design and scoring. Assessing grammar or vocabulary through decontextualised items risks construct under-representation, as it fails to capture how these resources are deployed in context and may introduce construct-irrelevant variance through test-taking strategies or format familiarity (Weir, 2005; Purpura, 2004). In contrast, embedding linguistic competence within tasks allows performance to be evaluated in relation to communicative effectiveness.

Accordingly, in the NEAS test, grammatical control is assessed through the use and interpretation of structure across all four skills; lexical competence is evaluated through appropriacy, precision, and contextual interpretation; discourse competence is reflected in the organisation and cohesion of extended meaning; and pragmatic competence is realised through task responsiveness, tone, and appropriacy to communicative purpose.

This approach ensures that linguistic features are interpreted in relation to meaning rather than treated as ends in themselves. It also enhances ecological validity, as it reflects the integrated nature of language use outside the testing context (Bachman & Palmer, 1996; Weir, 2005).

3.3. Progression, difficulty, and level differentiation

Across all four skills, proficiency is treated as a progression from concrete, literal, and highly supported performance at lower levels to more abstract, discourse-based, and evaluative performance at higher levels. At lower levels, candidates are required to identify, convey, or respond to essential meaning in familiar contexts. At intermediate levels, they are increasingly required to organise information, integrate ideas, provide reasons or explanations, and manage connected discourse. At advanced levels, they must handle abstraction, qualification, stance, and more flexible discourse management across receptive and productive tasks. This progression reflects the CEFR's conceptualisation of proficiency as qualitative change in control, complexity, and communicative capacity (Council of Europe, 2001, 2020).

A central implication of this model is that difficulty is determined by processing and communicative demands rather than by surface features alone. Across the test, higher-level performance is associated with increasing control over inference, integration, discourse organisation, abstraction, and communicative purpose. Surface features such as sentence length, word frequency, or response length may provide useful diagnostic information, but they do not determine level by themselves (Weir, 2005; Crossley, Greenfield, & McNamara, 2008).

This principle applies across both receptive and productive skills. In reading and listening, difficulty arises from the cognitive demands of meaning construction, including inferencing, integration across clauses or stretches of discourse, and interpretation of abstract or qualified meaning (Alderson, 2000; Grabe, 2009; Khalifa & Weir, 2009; Buck, 2001; Field, 2008). In speaking and writing, difficulty arises from the degree of control, development, organisation, and abstraction demonstrated within the task (Bachman & Palmer, 1996; Weigle, 2002; Fulcher, 2003). Across all skills, level distinctions are therefore grounded in what candidates are able to do with language rather than in isolated formal features.

3.4. Test use cases and implication for design

The NEAS Test of English is designed to support multiple use cases, including placement, level certification, and benchmarking. These purposes impose different, and sometimes competing, requirements on test design. Placement decisions require efficient

classification of candidates into appropriate instructional levels; level testing focuses on confirming whether a candidate has achieved a particular standard; and benchmarking requires comprehensive coverage of the construct in order to support fine-grained interpretation of proficiency across levels.

Research in language testing indicates that classification accuracy is strengthened when assessment samples performance across a range of difficulty levels rather than relying on a single level-specific task (Bachman & Palmer, 1996; Fulcher, 2010). Restricting assessment to a narrow band of difficulty increases the risk of misclassification, particularly through ceiling and floor effects, where tasks are either too easy or too difficult to discriminate effectively between adjacent proficiency levels (Green, 2014; Weir, 2005). For this reason, the NEAS test adopts a full-range design in its receptive components, in which candidates engage with tasks spanning multiple CEFR levels. This allows level to emerge from patterns of performance across a range of processing demands, rather than from success on a single level-specific task.

This approach is particularly important for benchmarking purposes, where the aim is not only to assign a level but to characterise performance across the construct. Broad sampling enhances score interpretability and supports more robust alignment with CEFR descriptors, which are themselves defined as developmental continua rather than discrete thresholds (Council of Europe, 2001, 2020). At the same time, the design remains compatible with placement use by allowing lower-level candidates to demonstrate ability through successful engagement with less demanding tasks, while higher-level candidates differentiate themselves through performance on more complex tasks.

Taken together, this multi-purpose design reflects the principle that valid language assessment must balance efficiency, comparability, and construct coverage. By sampling performance across a structured range of difficulty, the NEAS test reduces misclassification risk, supports meaningful level decisions, and provides interpretable evidence of proficiency across CEFR bands.

3.5. Topic selection, schema activation, and cultural fairness

Across the test as a whole, topic selection is constrained by the need to activate generalisable schemas while avoiding dependence on culturally specific, identity-linked, or specialist knowledge. Tasks and texts are therefore based on domains such as everyday systems, procedures, study and work practices, common social situations, and broad social or organisational issues. This approach reflects research demonstrating that background knowledge and schema familiarity can significantly influence comprehension and performance independently of language ability (Alderson, 2000; Grabe, 2009; Rost, 2011).

This principle is central to construct validity and fairness. Performance should depend primarily on language ability rather than prior familiarity with a particular cultural domain, academic tradition, or specialist topic. Accordingly, the test avoids prompts and texts that rely heavily on sport, popular culture, national history, literary interpretation, or professional expertise. Instead, it privileges broadly accessible domains that support

meaningful language use across diverse candidate populations, thereby reducing construct-irrelevant variance (Bachman & Palmer, 1996; Weir, 2005; McNamara, 1996; Weigle, 2002; Fulcher, 2003).

This approach also facilitates the creation of parallel forms, since general topic frameworks can be replicated across administrations while maintaining stable discourse demands and construct coverage (Weir, 2005).

3.6. Task design, comparability, and observability

A further design principle is that tasks must permit valid comparison across candidates and support scoring based on observable performance. For this reason, the test prioritises controlled, clearly specified tasks that elicit construct-relevant performance without relying on hidden expectations or incidental difficulty (Bachman & Palmer, 1996; Weir, 2005).

Across the test, difficulty is engineered through the communicative and cognitive demands of the task rather than through obscure prompts, weighted scoring, or unnecessary rhetorical complexity. Where common task sets are used, as in speaking and writing, level distinctions emerge through the quality, complexity, and communicative effectiveness of performance within shared tasks rather than through differential task assignment (Weigle, 2002; Fulcher, 2003). Where item-based formats are used, as in reading and listening, different item types are selected to sample distinct aspects of the construct while maintaining objective scoring and minimising scorer variability (Alderson, 2000).

This emphasis on comparability is closely tied to observability. Performance must be describable in terms of features that can be reliably identified in candidate output, whether by human raters or automated systems. This principle underpins both task construction and the later development of scoring criteria, particularly in AI-mediated assessment contexts (Chapelle & Chung, 2010; Xi, 2010).

3.7. AI-mediated constraints and design implications

The design of the NEAS Test of English is further shaped by its delivery in an AI-mediated, unsupervised environment. This introduces constraints that influence both construct representation and task design.

A primary requirement is the standardisation of input and comparability of performance across candidates. In contrast to face-to-face assessments, where interlocutor behaviour may vary, the use of fixed prompts and structured task formats ensures that all candidates respond to the same stimuli. While this limits the inclusion of spontaneous interaction, it reduces construct-irrelevant variance and enhances reliability in automated contexts (Bachman & Palmer, 1996; Xi, 2010; Chapelle & Chung, 2010).

A second consideration is the need for observable and measurable performance features. Automated scoring systems rely on indicators such as response length, pausing patterns, lexical diversity, syntactic structure, and the use of cohesive devices (Chapelle & Chung,

2010; Xi, 2010). As a result, scoring criteria are framed in terms of identifiable behaviours. This does not reduce the construct to surface features alone; rather, it requires that higher-level constructs such as fluency, coherence, and task responsiveness be operationalised through observable indicators.

These indicators are interpreted in relation to communicative effectiveness. Hesitation, for example, is relevant only insofar as it affects intelligibility or coherence, and linguistic errors are evaluated in terms of their impact on meaning. This maintains alignment with communicative models of language ability while supporting reliable measurement.

The AI-mediated context necessarily limits the inclusion of interactional negotiation as part of the construct. While interactional competence is an important aspect of language use, it is difficult to assess reliably in automated environments due to variability in turn-taking and interlocutor behaviour (Galaczi, 2014). The speaking component therefore focuses on elicited monologic and semi-structured responses, which allow for the assessment of discourse management, fluency, and pragmatic appropriacy under controlled conditions.

Finally, the requirement for consistency across automated and human scoring contexts has informed the development of analytic scoring criteria. Descriptors are behaviourally anchored, clearly defined, and aligned with measurable features, ensuring compatibility with both human raters and AI-based scoring systems (Weir, 2005).

3.8. Summary of construct and design principles

Taken together, the design of the NEAS Test of English reflects a coherent application of the theoretical framework outlined in Section 2. The test is structured around the four macro-skills as the primary means of eliciting communicative performance. Linguistic competence is embedded within task execution, ensuring that grammar, vocabulary, discourse, and pragmatics are assessed in relation to meaning. Progression across levels is defined by increasing demands in processing, discourse management, abstraction, and communicative control rather than by surface features alone. Topic selection, task design, and scoring all prioritise fairness, comparability, and construct relevance. At the same time, the design accommodates the constraints of AI-mediated delivery by prioritising standardisation, observability, and reliability while maintaining alignment with established models of language proficiency and CEFR descriptors.

4. Receptive Skills (Reading & Listening)

4.1. Receptive Skills Overview

The receptive components of the NEAS Test of English assess candidates' ability to construct meaning from written and spoken input across progressively more demanding levels of discourse, inference, and abstraction. As outlined in Section 3, receptive proficiency is understood as a progression from concrete, explicit, and highly supported comprehension at lower levels to more abstract, discourse-based, and evaluative interpretation at higher levels (Council of Europe, 2001, 2020; Khalifa & Weir, 2009). This

construct reflects established models of reading and listening as processes involving the integration of local decoding, global comprehension, inferencing, and discourse-level meaning construction (Alderson, 2000; Buck, 2001; Grabe, 2009; Vandergrift & Goh, 2012).

Within this shared construct, reading and listening differ not in their overall interpretive purpose, but in the specific conditions under which meaning must be processed. Reading allows candidates to control pace, revisit earlier input, and manage attention visually across the text. Listening, by contrast, is transient and time-bound, requiring candidates to process meaning as input unfolds (Buck, 2001; Field, 2008; Rost, 2011).

For this reason, the two receptive skills are aligned at the level of construct, progression, fairness, and scoring logic, but differ in the design parameters through which difficulty is realised. Reading difficulty is shaped primarily through text and task characteristics, whereas listening difficulty depends additionally on delivery conditions such as speech rate, pausing, text length in time, repetition, and accent (Buck, 2001; Field, 2008; Weir, 2005).

The sections below therefore focus on the skill-specific realisation of the receptive construct in reading and listening.

4.2. Reading-Specific Design Parameters

4.2.1. Text design principles

The reading test consists of multiple text–task sets distributed across CEFR levels (A2, B1, B2, and C1), with two task sets at each level. All candidates complete the full set.

This design reflects the principle that receptive proficiency is best interpreted through performance across a range of processing demands rather than through performance on a single level-specific task. Construct-based models of reading emphasise that comprehension ability emerges through the interaction between reader resources and varying text and task demands, rather than as a fixed response to a single level of input (Alderson, 2000; Khalifa & Weir, 2009). Lower-level candidates may engage primarily with lower-level texts, while higher-level candidates demonstrate their ability across the full range, including more demanding discourse (Bachman & Palmer, 1996; Weir, 2005).

Difficulty is therefore not adapted to the candidate in advance but emerges from the interaction between candidate ability and the full range of tasks presented. This approach aligns with criterion-referenced assessment principles, in which performance is interpreted against a defined continuum of proficiency rather than adjusted dynamically through test adaptation (Bachman & Palmer, 1996; Weir, 2005).

Reading texts are written specifically for assessment purposes, with the aim of ensuring that difficulty arises from the intended construct rather than from incidental

stylistic features. Unnecessary rhetorical density and stylistic complexity are avoided in order to minimise construct-irrelevant variance (Alderson, 2000; Weir, 2005).

Difficulty is calibrated through interacting parameters including conceptual density, lexical abstraction, discourse organisation, and inferential demand. Research in reading demonstrates that comprehension difficulty is driven primarily by the need to integrate information, construct inferences, and process discourse structure, rather than by surface features such as sentence length or word frequency alone (Grabe, 2009; Khalifa & Weir, 2009; Crossley et al., 2008). At lower levels, texts are shorter, more concrete, and explicitly organised. At higher levels, the key differentiators are abstraction, discourse structure, and the need to integrate meaning across larger stretches of text.

At A2, texts focus on routine, concrete content with explicit sequencing and minimal inference. At B1, texts introduce basic explanation and limited inferencing within familiar domains. At B2, texts require integration across paragraphs, discourse awareness, and lexical inference within moderately abstract contexts. At C1, texts involve abstract, evaluative, and discourse-rich content requiring interpretation of stance, qualification, and reasoning, consistent with CEFR descriptors for advanced reading proficiency (Council of Europe, 2001, 2020).

4.2.2. Lexical profiling and text calibration

Lexical profiling is used as a diagnostic tool during text development, not as a determinant of level. Automated metrics can identify anomalies and support consistency, but they do not reliably represent processing demands, particularly in short or genre-specific texts (Crossley et al., 2008).

Text calibration therefore depends on interpreting lexical data in relation to discourse purpose, conceptual accessibility, and task demands. Apparent mismatches between profile labels and intended levels are resolved through analysis of processing load rather than reliance on frequency-based measures alone.

4.2.3. Text length, timing, and processing load

The reading test is designed to be completed within approximately 60–72 minutes, with a total administration time of around 75 minutes once instructions and transitions are included. This timing model was determined not simply by overall text length, but by the number of text–task sets in the test, the number of questions attached to each set, and the level of cognitive processing required by those questions. Lower-level tasks require less time because they focus more heavily on explicit information and basic gist, whereas higher-level tasks require more time for inferencing, discourse integration, lexical interpretation, and the processing of abstract or qualified meaning.

The allocation of additional time at higher levels is therefore based on the interaction between task quantity and task complexity rather than on text length alone. This

approach is consistent with research showing that reading tasks involving inferencing, integration across paragraphs, and discourse-level comprehension require more time than tasks focused on expeditious reading or explicit detail, even where texts are not dramatically longer (Alderson, 2000; Grabe, 2009; Khalifa & Weir, 2009; Weir, 2005). Text length is accordingly treated as a contributing factor, but not as the sole or primary determinant of timing. This avoids reducing advanced reading to a test of stamina while ensuring that candidates have sufficient time to engage in the intended reading processes.

4.2.4. Pilot task validation and text appropriateness

Pilot analysis confirmed that automated text metrics must be interpreted cautiously when evaluating reading difficulty. In several cases, lower-level texts attracted higher lexical profile classifications due to short-text effects, formulaic expressions, or genre-specific features, despite placing relatively low demands on inferencing and discourse integration. Conversely, some higher-level texts exhibited increased lexical density without exceeding the intended construct level, as their primary difficulty arose from discourse organisation and conceptual complexity rather than vocabulary alone.

This finding is consistent with research demonstrating that surface-level metrics such as word frequency, lexical density, or readability scores do not reliably predict comprehension difficulty, particularly in shorter texts or controlled assessment contexts (Crossley, Greenfield, & McNamara, 2008; Alderson, 2000). Reading difficulty is instead determined by the interaction of lexical, syntactic, discourse, and conceptual factors, including the need to integrate information across clauses and paragraphs and to construct inferences (Grabe, 2009; Khalifa & Weir, 2009).

Validation decisions were therefore based on a multi-factor analysis of processing demands rather than on lexical metrics alone, ensuring alignment with the intended proficiency levels and the underlying construct of reading comprehension (Weir, 2005).

4.3. Listening-Specific Design Parameters

As with the reading test, the listening test consists of multiple text–task sets distributed across CEFR levels (A2, B1, B2, and C1), with two task sets at each level. All candidates complete the full set.

This design reflects the principle that listening proficiency is best interpreted through performance across a range of processing conditions rather than through a single level-specific task. Because listening involves additional constraints such as transience and real-time processing, sampling performance across levels allows more reliable differentiation of candidates' ability to manage increasing cognitive and perceptual demands (Buck, 2001; Field, 2008; Weir, 2005).

4.3.1. Real-time processing and delivery constraints

Listening differs from reading in that it is time-bound and transient. Candidates must process input as it unfolds, without the ability to revisit earlier material. For this reason, difficulty arises from the interaction of processing demands and delivery conditions rather than from linguistic content alone (Buck, 2001; Field, 2008; Vandergrift & Goh, 2012). Key parameters such as speech rate, pausing, text length, and number of listens are therefore treated as core construct variables.

4.3.2. Control of speech rate, delivery, and pausing

Speech rate is controlled to ensure that comprehension rather than perceptual decoding speed is being assessed. Research consistently demonstrates that increased speech rate reduces comprehension, particularly for lower-proficiency listeners, while slower and more clearly articulated speech supports segmentation and lexical access (Buck, 2001; Field, 2008; Griffiths, 1992; Zhao, 1997).

Approximate speech rate ranges are therefore specified by level:

- A2: 110–130 words per minute
- B1: 130–150 words per minute
- B2: 150–170 words per minute
- C1: 170–190 words per minute

These ranges reflect a progression from carefully controlled delivery toward natural speech rates typical of real-world communication.

Pausing is also controlled as part of delivery. Lower-level texts include slightly extended pauses at clause boundaries to support processing, while higher-level texts rely on natural discourse pausing. This reflects findings that pausing can facilitate comprehension by reducing processing load without fundamentally altering discourse structure (Field, 2008).

4.3.3. Control of text length and cognitive load

Listening text length is specified in time rather than words, reflecting the constraints of working memory and attention in processing spoken input (Buck, 2001; Rost, 2011). As text length increases, listeners must retain and integrate information across longer stretches of discourse, increasing cognitive load and processing difficulty (Field, 2008).

Approximate length ranges are therefore defined by level:

- A2: 45–60 seconds
- B1: 60–90 seconds
- B2: 90–120 seconds
- C1: 120–180 seconds

These ranges are designed to allow the intended comprehension processes to occur without overloading working memory. Difficulty is determined by the interaction of length, speech rate, and task demands rather than by duration alone.

4.3.4. Number of listens as a construct variable

The number of listening opportunities is treated as a construct parameter rather than an administrative decision. Research shows that repetition fundamentally alters listening processing: a single listen emphasises real-time comprehension, while multiple listens allow monitoring, confirmation, and repair (Buck, 2001; Field, 2008; Vandergrift & Goh, 2012).

The test therefore adopts the following model:

- A2–B1: Two plays
- B2: One or two plays (transitional)
- C1: One play

Multiple listens at lower levels support the transition from perceptual decoding to meaning construction, while reduced repetition at higher levels reflects the expectation of sustained real-time comprehension.

4.3.5. Accent control and authenticity

Accent variation is incorporated in a controlled manner to balance authenticity and fairness. Research shows that unfamiliar accents can affect comprehension independently of language proficiency and may introduce construct-irrelevant variance if not managed systematically (Harding, 2012; Field, 2008).

The test therefore adopts a staged approach:

- A2: Single, standard accent
- B1: Limited variation
- B2: Controlled variation
- C1: Wider accent range

At lower levels, accent is controlled to minimise phonological novelty. At higher levels, variation reflects real-world diversity while avoiding excessive interaction with other sources of difficulty.

4.3.6. Administration model and method effects

The listening test uses a structured administration model with controlled preview time, listening phases, and response opportunities. Previewing supports schema activation and reduces working-memory load but is standardised to avoid altering task demands (Sherman, 1997; Chang & Reed, 2006).

Questions remain visible during listening to support integration of information rather than recall, minimising construct-irrelevant memory demands.

4.3.7. Role of lexical profiling in listening validation

Lexical profiling is used diagnostically but is not determinative. Because profiling tools are designed for written text, they tend to overestimate the difficulty of spoken input by ignoring pausing, redundancy, and discourse predictability (Field, 2008; Buck, 2001).

Final classification is therefore based on listening-specific parameters, including speech rate, text length, repetition, discourse structure, and task demands.

4.3.8. Summary of Receptive Skills Design Rationale

The receptive components of the NEAS Test of English are grounded in a unified construct of meaning construction across written and spoken input. Within this shared construct, reading and listening differ in the conditions under which comprehension takes place and therefore in the design parameters used to calibrate difficulty. Reading-specific design focuses on text structure, lexical calibration, and cognitive load in relation to visually accessible input, while listening-specific design additionally requires control of speech rate, pausing, repetition, accent, and administration conditions. Together, these elements support valid and reliable measurement of receptive language proficiency across CEFR levels.

5. Productive Skills (Speaking & Writing)

5.1. Productive Skills Overview

The productive components of the NEAS Test of English assess candidates' ability to produce intelligible, coherent, and contextually appropriate language in written and spoken modes across progressively more demanding levels of discourse, development, and abstraction. As set out in Section 3, productive proficiency is understood as a progression from simple, concrete, and highly supported communication at lower levels to more extended, structured, and abstract discourse at higher levels (Council of Europe, 2001, 2020; Weir, 2005). Within this shared construct, both writing and speaking require candidates to manage communicative purpose, discourse organisation, lexical and grammatical control, and increasing levels of abstraction, consistent with contemporary models of communicative language ability (Bachman, 1990; Weigle, 2002; Fulcher, 2003).

The two productive skills, however, differ in the conditions under which this performance is realised. Writing permits planning, drafting, monitoring, and revision, whereas speaking requires real-time production under fixed temporal conditions. This distinction is widely recognised in language assessment research, which characterises writing as a recursive and revisable process and speaking as time-constrained performance requiring online processing and immediate formulation (Weigle, 2002; Levelt, 1989; Fulcher, 2003).

Both skills therefore share a common construct architecture but differ in the parameters through which task demands, processing load, and standardisation are implemented. The sections below focus on the skill-specific realisation of the productive construct in writing and speaking.

5.2. Writing-Specific Design Parameters

5.2.1. Task design and comparability

The writing test consists of two tasks completed by all candidates: a short functional message and a short discursive response. These tasks sample two central forms of real-world writing—transactional communication and short explanatory or discursive writing—while remaining sufficiently controlled for reliable scoring (Weigle, 2002; Weir, 2005).

Both tasks are scalable across CEFR levels. Lower-level candidates can communicate essential information in simple terms, while higher-level candidates can demonstrate greater control through fuller development, clearer structure, and more precise language. Writing ability is therefore differentiated through performance within shared tasks rather than through task assignment, consistent with research emphasising the importance of common prompts for comparability in writing assessment (Weigle, 2002; Weir, 2005).

5.2.2. Discourse organisation and idea development

Writing is treated as purposeful, discourse-level communication. Candidates are required not only to produce language, but to address the task, organise ideas coherently, and support those ideas where appropriate. This reflects the view that writing proficiency involves the ability to manage discourse, develop ideas, and achieve communicative purpose, rather than simply produce grammatically correct sentences (Weigle, 2002; Hyland, 2004; Weir, 2005).

Difficulty is calibrated through the extent of organisation, development, and abstraction rather than through length alone. From approximately B1 upward, the ability to structure and extend a response becomes a more reliable indicator of proficiency than sentence-level accuracy, consistent with CEFR descriptors and research showing that discourse control and development become increasingly central at higher proficiency levels (Council of Europe, 2001, 2020; Weigle, 2002).

5.2.3. Lexical profiling and task calibration

Lexical profiling is used diagnostically during task development but is not determinative. Research has shown that surface-level textual measures, including lexical frequency and readability indices, do not reliably capture writing difficulty, which depends on communicative purpose, conceptual scope, and discourse demands (Crossley, Greenfield, & McNamara, 2008; Weigle, 2002).

Task classification is therefore based on CEFR alignment and discourse requirements rather than global textual metrics alone, ensuring that difficulty reflects the underlying construct of writing ability rather than superficial textual features (Weir, 2005).

5.2.4. Timing, planning, and revision

The writing test allows approximately 50–60 minutes, with both tasks visible from the outset. This supports planning, drafting, and revision as integral components of writing performance. Research in writing assessment consistently characterises writing as a recursive process involving planning, formulation, monitoring, and revision, rather than one-pass production (Flower & Hayes, 1981; Weigle, 2002).

Because writing is not transient, the opportunity to revise is treated as part of the construct rather than a distortion. Limiting revision opportunities would risk shifting the construct toward speed of production rather than communicative effectiveness (Weigle, 2002; Weir, 2005).

5.2.5. Pilot validation

Validation confirmed that the tasks function appropriately across proficiency levels. Lower-level candidates can respond meaningfully, while higher-level candidates can demonstrate increased development, organisation, and control. This supports the effectiveness of the fixed-task design and is consistent with research indicating that well-designed common tasks can elicit performance across a wide proficiency range when they are sufficiently open and scalable (Weigle, 2002; Weir, 2005).

5.3. Speaking-Specific Design Parameters

5.3.1. Task design and structure

The speaking test consists of four parts: short responses to familiar questions, a short long-turn task, two related discussion questions, and a situational role-play response. These tasks sample complementary aspects of spoken performance, including short responses, sustained turns, explanation, and pragmatic language use. This multi-part structure reflects established approaches in speaking assessment, where different task types are used to elicit a range of discourse functions and interactional demands within a single test (Fulcher, 2003; Weir, 2005; O'Sullivan, 2011).

All tasks are scalable across CEFR levels, allowing differentiation through performance rather than task assignment. Research in speaking assessment demonstrates that common tasks can validly elicit performance across a wide proficiency range when they are sufficiently open and allow for variation in length, complexity, and discourse development (Fulcher, 2003; Weir, 2005).

5.3.2. Discourse management and real-time processing

Speaking is assessed as discourse performance under real-time conditions. Candidates must produce intelligible and coherent responses while managing time pressure, fluency, and idea development. This reflects psycholinguistic models of speech production, which conceptualise speaking as an online process involving conceptualisation, formulation, and articulation under temporal constraints (Levelt, 1989), as well as applied models of speaking proficiency that emphasise fluency, coherence, and discourse management (Fulcher, 2003).

Difficulty is calibrated through the ability to sustain speech, organise ideas, express reasons, and manage increasing abstraction. From approximately B1 upward, the ability to maintain a turn and develop a response becomes a key indicator of proficiency, consistent with CEFR descriptors and research showing that extended discourse and explanation are central features of intermediate and advanced speaking ability (Council of Europe, 2001, 2020; Fulcher, 2003).

5.3.3. Delivery conditions and standardisation

The speaking test is delivered in an AI-mediated, unsupervised format with controlled timing and standardised prompts. Candidates respond within fixed preparation and response windows, and no live interlocutor support is provided.

This design reflects research on technology-mediated speaking assessment, which emphasises the importance of standardisation in input, timing, and task delivery to ensure comparability of performance across candidates (Xi, 2010; Chapelle & Chung, 2010). By removing variability associated with interlocutor behaviour, the test reduces construct-irrelevant variance and enhances reliability.

Features such as spontaneous negotiation of meaning are not included, as they are difficult to standardise and may introduce variability related to interactional support rather than candidate ability. This is consistent with research showing that interlocutor effects can significantly influence speaking performance and complicate score interpretation (Galaczi, 2014; Xi, 2010).

5.3.4. Sequencing and cognitive load

The task sequence moves from short, familiar responses to more extended and demanding tasks. This progression supports candidate engagement while allowing systematic sampling of sustained production, reasoning, and pragmatic control.

Research in speaking assessment suggests that staged task sequences can reduce initial anxiety, establish baseline performance, and progressively elicit more complex discourse, leading to more representative samples of speaking ability (Fulcher, 2003; O'Sullivan, 2011). The progression from simple to more demanding tasks also aligns with cognitive models of speech production, in which increased planning and formulation demands are associated with longer and more complex turns (Levelt, 1989).

5.3.5. Pilot validation

Validation confirmed that all four tasks function across proficiency levels. Lower-level candidates can produce simple, relevant responses, while higher-level candidates can demonstrate extended discourse, greater control, and increased abstraction.

This supports the effectiveness of the fixed-task design and is consistent with research indicating that well-designed speaking tasks can elicit performance across a wide proficiency range when they allow for variation in response length, complexity, and discourse development (Fulcher, 2003; Weir, 2005).

5.4. Summary of Productive Skills Design Rationale

The productive components of the NEAS Test of English are grounded in a unified construct of communicative language production across spoken and written modes. Within this shared construct, writing and speaking differ in the processing conditions under which performance is produced and therefore in the design parameters through which difficulty is calibrated. Writing-specific design focuses on planning, revision, discourse development, and task comparability in a non-transient mode, while speaking-specific design focuses on real-time production, temporal control, staged elicitation, and standardisation in an AI-mediated environment. Together, these elements support valid and reliable measurement of productive language proficiency across CEFR levels.

6. Assessment and Scoring Framework

6.1. Scoring construct and underlying principles

The scoring framework of the NEAS Test of English is designed to ensure that reported scores represent meaningful differences in language proficiency, aligned to CEFR levels and grounded in observable performance. Across all four skills, scoring is based on the principle that language ability is best understood as controlled performance within communicative tasks rather than as the accumulation of discrete linguistic features or isolated errors (Bachman & Palmer, 1996; Weir, 2005).

A central principle underpinning the framework is that CEFR does not require perfection but control appropriate to level. Performance is therefore evaluated in terms of the extent to which language resources are controlled sufficiently to achieve communicative purposes. Errors are interpreted functionally, in relation to their frequency, distribution, and impact on meaning, rather than as deviations from an idealised native-speaker norm (Council of Europe, 2001, 2020).

Across all components, the scoring model reflects three core principles: communicative effectiveness over formal accuracy alone, observable performance rather than inferred ability, and progression through increasing control, complexity, and impact on meaning. These principles apply across both productive and receptive skills, although they are operationalised differently according to the nature of the construct.

6.2. Analytic scoring in productive skills

In speaking and writing, performance is evaluated using analytic rating scales in which separate but related dimensions of language ability are assessed independently. This approach reflects established practice in large-scale language testing and is supported by research showing that analytic scoring improves diagnostic value and supports greater rating consistency (Weigle, 2002; Fulcher, 2003).

6.2.1. Rationale for analytic criteria

The use of multiple criteria is based on the understanding that language ability is multi-dimensional and that different aspects of performance develop at different rates (Larsen-Freeman, 2006). A single holistic score risks masking important differences in performance, whereas separate criteria allow construct-relevant distinctions to be captured more accurately.

At the same time, criteria are defined to minimise overlap. Task fulfilment evaluates relevance and completeness of response rather than linguistic quality; cohesion and organisation operationalise discourse control through observable features such as linking and information flow; grammar and vocabulary are assessed in terms of control and impact on meaning rather than formal correctness alone; and, in speaking, fluency and cohesion are combined where separation would reduce reliability and introduce artificial distinctions. The resulting framework reflects a necessary balance: criteria must be distinct enough to be meaningful but integrated enough to reflect actual language use (Bachman, 1990; Fulcher, 2003).

6.3. Comprehension-based scaling and error interpretation

Across productive-skill criteria, performance is differentiated primarily by the impact of language use on comprehension. At higher levels, errors are rare or do not affect comprehension; at mid levels, they may cause local difficulty but meaning remains recoverable; at lower levels, they significantly reduce or prevent understanding.

Research in both speaking and writing assessment demonstrates that error frequency alone is an unreliable indicator of proficiency, whereas error gravity provides a more valid basis for distinguishing levels (Alderson et al., 1995; Weigle, 2002). The framework also recognises the accuracy–complexity trade-off: higher-level candidates may produce more errors because they attempt more complex structures or express more nuanced ideas. Such errors are not penalised if overall meaning remains clear, consistent with findings in second language acquisition research (Ellis, 1994; Larsen-Freeman, 2006).

6.4. Scoring in receptive skills

In reading and listening, scoring follows a dichotomous model in which each item is marked as correct or incorrect, with no partial credit awarded. This reflects standard practice in receptive-skills assessment and the principle that comprehension is a latent construct inferred from patterns of successful interpretation across items rather than directly observable (Alderson, 2000).

6.4.1. Rationale for dichotomous scoring

Binary scoring is adopted because it supports objectivity, reliability, and construct clarity. Closed-response formats minimise scorer interpretation and ensure consistency across administrations. They also allow reading and listening to be assessed independently of productive ability, since requiring constructed responses would introduce writing or speaking as confounding factors. At a psychometric level, item-level scoring supports stable measurement across forms and facilitates calibration.

Importantly, difficulty is controlled through text complexity, processing demands, and task type rather than through weighted scoring. This ensures that difficulty arises from the construct being measured rather than from hidden scoring mechanics (Weir, 2005; Fulcher, 2010).

6.5. CEFR alignment and scale design

All scoring criteria and item-based scores are mapped to CEFR levels using a common scale structure, with sub-bands such as A2+, B1+, and B2+ used to capture intermediate performance. This reflects the functional nature of CEFR proficiency, in which levels represent what users can do consistently rather than isolated instances of higher-level performance, and the practical need for finer distinctions in placement and reporting (Council of Europe, 2020).

The use of a unified 10–1 scale across criteria supports consistency in interpretation, transparency in reporting, and alignment across skills. CEFR alignment is treated as an interpretive framework rather than a mechanical conversion: scores are linked to CEFR levels through descriptor alignment, task design, and scoring logic rather than through statistical equating alone (Council of Europe, 2001; Weir, 2005).

6.6. Combining criteria: decision rules and overall level assignment

Because productive skills are assessed analytically, a principled method is required to combine criterion-level scores into an overall CEFR level. The NEAS test adopts a profile-based decision model in which a level is awarded when most criteria demonstrate performance at that level and major weaknesses prevent upward classification.

This reflects the principle that language proficiency is multi-dimensional but integrated. A candidate cannot be considered to operate fully at a given level if one core component is substantially weaker, even where other areas are strong.

6.6.1. Rationale for decision rules

The use of explicit decision rules is supported by language testing theory, which emphasises the need for transparent and interpretable score meaning (Bachman & Palmer, 1996; Council of Europe linking manuals). Such rules ensure that no single criterion dominates disproportionately, that strengths do not mask critical weaknesses, and that decisions remain consistent and reproducible. They also reflect

the logic of real-world language use, where effective communication depends on balanced control across multiple dimensions rather than isolated strengths.

6.7. Consistency across skills

Although scoring mechanisms differ between productive and receptive skills, the overall framework is unified by a common set of principles. Productive skills are scored through analytic criteria, while receptive skills use dichotomous item scoring; productive difficulty is expressed through quality of performance, while receptive difficulty is controlled through text and task demands; and across both, performance is interpreted in relation to communicative effectiveness, control, and level-appropriate functioning. This ensures that all components contribute to a coherent interpretation of proficiency rather than functioning as separate or incompatible measures.

6.8. Summary of scoring rationale

Taken together, the assessment and scoring framework is grounded in a coherent view of language proficiency as controlled, communicative performance across skills. Productive skills are assessed analytically to capture multiple dimensions of language use, while receptive skills are assessed through objective item-based scoring. Across all components, performance is interpreted in relation to communicative effectiveness, control, and impact on meaning rather than formal accuracy alone.

CEFR alignment is achieved through descriptor-based interpretation, supported by transparent decision rules and a consistent scale design. The resulting framework ensures that scores are meaningful, comparable, and defensible, and that they reflect genuine differences in language ability rather than artefacts of task design or scoring procedure.

6.9. Rubric and Descriptor Design Principles

6.9.1. Conceptualising descriptors as measurement instruments

The rating scales used in the NEAS Test of English are treated not as informal guidance for raters, but as central measurement instruments within the assessment system. In performance-based language assessment, descriptors operationalise the construct by defining what counts as evidence of ability and how differences in performance are interpreted (Bachman & Palmer, 1996; Fulcher, 2010). Descriptor design is therefore approached as a core component of test validity rather than as a secondary or cosmetic feature.

Descriptors are written to satisfy three simultaneous requirements: they must represent the intended construct, describe features that can be reliably identified in performance, and support consistent differentiation across levels. These requirements inform all aspects of rubric design across speaking and writing.

6.9.2. Observable behaviour and avoidance of abstract constructs

A key design principle is that descriptors are grounded in observable linguistic behaviour rather than abstract or inferential qualities. Terms such as coherence, fluency, or control are operationalised through identifiable features such as the use and accuracy of cohesive devices, clarity of pronominal reference, hesitation patterns, response relevance, and the impact of errors on comprehension.

This approach reflects a well-established concern in language testing: constructs that are not directly observable tend to reduce rating reliability and increase subjectivity (Weigle, 2002; McNamara & Roever, 2006). By anchoring descriptors in identifiable performance features, the rubric reduces ambiguity and supports both human and automated scoring.

6.9.3. Functional interpretation of error and control

Across all criteria, descriptors are structured around a functional interpretation of language use in which errors are evaluated in terms of their impact on meaning rather than their mere presence. This reflects the CEFR's emphasis on control appropriate to level rather than formal correctness or native-like performance (Council of Europe, 2001, 2020).

The scales therefore differentiate performance according to whether errors do not affect comprehension, occasionally affect clarity while meaning remains recoverable, frequently disrupt understanding, or prevent meaningful interpretation altogether. This progression allows tolerance of developmental errors at higher levels, recognition of successful communication despite imperfect form, and avoidance of penalising risk-taking in more complex language use. It also aligns with research showing that accuracy, complexity, and fluency develop unevenly and interactively (Larsen-Freeman, 2006).

6.9.4. Integration and separation of criteria

A critical aspect of rubric design is deciding which dimensions of performance should be assessed together and which should be separated. The NEAS rubric adopts a selective analytic approach in which distinct constructs are separated where they can be reliably distinguished, while interdependent constructs are integrated where separation would reduce reliability or create artificial distinctions.

In speaking, for example, fluency, cohesion, and task responsiveness are treated as a single criterion because they are co-constructed in real-time speech and difficult to separate reliably. In writing, task fulfilment is separated from linguistic criteria so that candidates are not penalised for language limitations when they nevertheless demonstrate clear task understanding. This approach reflects findings that over-segmentation can lead to double marking, construct overlap, and reduced inter-rater reliability (Fulcher, 2003; Weigle, 2002).

6.9.5. Developmental progression and scale logic

Descriptors are organised along a developmental continuum reflecting increasing control, flexibility, and communicative effectiveness. Rather than relying on impressionistic labels, the scales are built around systematic changes in range and flexibility of language use, consistency of control, ability to manage discourse and information flow, and the degree of listener or reader effort required.

At lower levels, descriptors emphasise limited control, frequent breakdowns in meaning, and reliance on simple or fragmented language. At higher levels, they emphasise consistent control, flexibility, and effective management of discourse and nuance. Progression is therefore defined not by the absence of error, but by increasing ability to maintain meaning despite linguistic limitations.

6.9.6. Avoidance of native-speaker norms and academic bias

Descriptors are explicitly designed to avoid two common sources of construct-irrelevant bias: native-speaker benchmarking and academic discourse bias.

6.9.7. Native-speaker benchmarking

The rubric does not require or reward native-like performance. Instead, it evaluates intelligibility, appropriacy, and communicative effectiveness. This reflects research showing that accent and non-native features do not necessarily impede communication and should not be treated as deficits unless they interfere with understanding (Munro & Derwing, 1995; Harding, 2013).

6.9.8. Academic discourse bias

Descriptors avoid assumptions tied to academic writing conventions unless these are explicitly required by the task. Instead, they focus on clarity of purpose, organisation of ideas, and appropriateness to communicative context. This ensures that the test remains valid across a wide range of candidates and does not privilege those with specific educational backgrounds (Hyland, 2004).

6.9.9. Compatibility with AI-mediated scoring

An additional design constraint is that descriptors must be interpretable not only by trained human raters, but also by AI-mediated scoring systems. Descriptors are therefore framed in terms of features that can be operationalised computationally, including response length and development, lexical diversity and error patterns, syntactic complexity and accuracy, pausing and fluency measures, and discourse markers and reference chains.

By avoiding vague or purely impressionistic language, the rubric supports scalable and consistent scoring across delivery modes without compromising construct validity (Xi, 2010; Chapelle & Chung, 2010).

6.9.10. Alignment with CEFR descriptors

The descriptor system is designed to align with the CEFR while avoiding direct replication of CEFR wording. This maintains interpretive compatibility with CEFR levels, supports finer distinctions through sub-bands, and preserves flexibility in task design and scoring.

Alignment is achieved through a shared emphasis on communicative effectiveness, a similar progression in control, range, and discourse management, and a consistent treatment of error as a function of level. This reflects recommended practice in CEFR-linked assessment, where alignment is based on functional equivalence rather than literal descriptor matching (Council of Europe, 2001; Weir, 2005).

6.9.11. Summary of descriptor design rationale

The rubric and descriptor system is grounded in a principled approach to performance assessment in which descriptors function as operational definitions of language ability. They are constructed to be observable, scalable, and aligned with communicative language use, while avoiding sources of bias and supporting both human and AI-mediated scoring.

By prioritising impact on meaning, integrating interdependent constructs, and reflecting developmental progression across levels, the descriptor system ensures that scores are interpretable, reliable, and valid representations of proficiency.

7. Scaling, CEFR Alignment, and Decision Rules

7.1. Conceptual basis for scaling and CEFR alignment

The NEAS Test of English adopts a criterion-referenced approach to scaling, in which performance is interpreted against clearly defined levels of language ability rather than relative to the performance of other candidates. In this framework, scores are intended to represent evidence of control across multiple dimensions of language use rather than numerical accumulation alone. The purpose of scaling is therefore to locate performance within a meaningful interpretive system that reflects communicative ability and supports transparent reporting (Council of Europe, 2001, 2020; Bachman & Palmer, 1996; Weir, 2005).

The test operationalises this principle by mapping analytic performance across criteria to CEFR bands through a structured decision process. This ensures that reported levels correspond to stable patterns of performance rather than isolated strengths or weaknesses.

7.2. Band structure and developmental scale

Performance in speaking and writing is first evaluated using analytic rating scales, with each criterion scored on a ten-point scale. These raw scores are then mapped onto CEFR-

aligned bands, including intermediate plus-levels such as A2+, B1+, B2+, and C1+, which provide finer discrimination within broad CEFR categories.

The mapping follows a monotonic progression in which higher scores correspond to greater control, flexibility, and communicative effectiveness:

Raw score	CEFR level				
10	C1+	6	B1+	2	A1
9	C1	5	B1	1	Below A1
8	B2+	4	A2+		
7	B2	3	A2		

This mapping reflects the functional interpretation of proficiency associated with the CEFR and provides greater granularity for placement and reporting purposes (Council of Europe, 2020; Green, 2014). Applying the same mapping across analytic criteria supports transparency and avoids hidden weighting.

7.3. Construct-based cut scores

Cut scores in the NEAS test are based on construct-based judgement anchored in descriptor performance rather than statistical norming alone. A given CEFR level is awarded only when a candidate demonstrates consistent control of the relevant abilities across multiple criteria, rather than occasional performance associated with that level.

This approach ensures that level decisions are tied directly to observable aspects of performance, such as maintenance of meaning, control over grammar, vocabulary, and discourse, and the degree of listener or reader effort required. Cut scores therefore function as interpretive thresholds grounded in the construct being assessed rather than as arbitrary numerical boundaries (Bachman & Palmer, 1996; Fulcher, 2010).

7.4. Profile-based decision rules

Overall CEFR levels are determined using a profile-based decision model rather than simple averaging. This reflects the multidimensional nature of language proficiency and avoids the risk of stronger performance in one area masking serious weakness in another.

The core principle is that a level is awarded only when there is sufficient consistency across criteria, combined with the absence of critical weaknesses. In practice, a candidate is assigned a CEFR level when the majority of criteria meet or exceed that level and no criterion falls substantially below it.

The model also incorporates plus-level decisions, ceiling constraints, and conservative treatment of borderline profiles. Plus-levels are awarded when performance consistently exceeds the base level across multiple criteria; ceiling constraints prevent the award of a high level where a major weakness is present; and borderline profiles are resolved

conservatively unless evidence for the higher level is clear. These rules ensure that level assignment remains stable, interpretable, and resistant to distortion by uneven performance (Weigle, 2002; Shaw & Weir, 2007).

7.5. Relationship between analytic profiles and overall level

The use of analytic scoring allows both criterion-level profiles and overall CEFR levels to be reported. This dual reporting model increases the interpretive value of the test by providing both a summary judgement of overall proficiency and diagnostic information about strengths and weaknesses across criteria.

The overall CEFR band is therefore not a simple arithmetic outcome, but a construct-based judgement derived from the performance profile. This supports both high-level reporting and more targeted interpretation for placement or development purposes (Bachman & Palmer, 1996; Weigle, 2002).

7.6. Alignment with external frameworks (IELTS and related scales)

For interpretive purposes, NEAS CEFR levels may be associated with approximate IELTS band ranges based on publicly available alignment research. These correspondences are intended to support external understanding and comparability, particularly in educational and migration contexts.

However, such mappings are approximate rather than equivalent. IELTS scores are derived from holistic banding procedures and test-specific constructs, whereas the NEAS test employs an analytic, profile-based model. Correspondence can therefore only be expressed in ranges rather than precise point matches (Taylor & Jones, 2006; Cambridge English, 2015). This distinction is maintained explicitly in reporting to avoid misinterpretation.

7.7. Validity and defensibility of the scaling model

The scaling and decision framework is defensible on both theoretical and practical grounds. It is construct-driven, transparent, consistent with CEFR principles, aligned with established assessment practice, and robust to uneven performance. By combining analytic scoring with structured decision rules, the model ensures that reported levels reflect communicative ability across multiple dimensions rather than artefacts of scoring or test design.

8. Validity and Reliability Considerations

8.1. Construct validity

The validity of the NEAS Test of English rests on the extent to which score interpretations accurately reflect the construct of communicative language ability across receptive and productive skills. Consistent with contemporary theory, validity is understood as an

evidence-based argument supporting the appropriateness and meaning of score use (Messick, 1989; Bachman & Palmer, 1996).

The test operationalises language ability as performance in communicative tasks requiring the integration of linguistic knowledge, cognitive processing, and discourse-level meaning construction. In reading and listening, this involves processes such as information location, inference, and discourse integration. In speaking and writing, it involves the production of intelligible, coherent, and contextually appropriate language.

Construct representation is supported by the alignment of tasks and scoring criteria with CEFR descriptors and by the use of task-based assessment rather than decontextualised item types. Difficulty is controlled through processing demands and discourse complexity rather than surface features alone, ensuring that progression reflects meaningful differences in communicative ability (Alderson, 2000; Weir, 2005).

Construct validity is further strengthened by the exclusion of construct-irrelevant factors. The test does not reward topic knowledge, cultural familiarity, or native-like performance where these are not part of the construct, ensuring that scores reflect language ability rather than extraneous influences.

8.2. Fairness and avoidance of construct-irrelevant variance

Fairness is addressed through systematic control of construct-irrelevant variance across all components of the test. Topics are selected to activate generalisable schemas while avoiding dependence on culturally specific or identity-linked knowledge. This reduces the influence of background knowledge on performance and supports comparability across diverse candidate groups (Grabe, 2009; Weir, 2005).

In productive skills, content is evaluated in terms of task fulfilment rather than topic sophistication, ensuring that candidates are assessed on language use rather than ideas or experience. In speaking, pronunciation is evaluated in terms of intelligibility and listener effort rather than conformity to native-speaker norms, supporting equitable assessment across linguistic backgrounds (Munro & Derwing, 1995).

The use of closed-response formats in receptive skills and analytic criteria in productive skills further contributes to fairness by reducing scorer subjectivity and ensuring consistent evaluation (Bachman & Palmer, 1996; Fulcher, 2010).

8.3. Reliability and consistency of measurement

Reliability is supported through a combination of task design, scoring methods, and standardisation procedures that ensure consistency of measurement across candidates and administrations (Weir, 2005).

In reading and listening, reliability is enhanced through objective, closed-response item types and dichotomous scoring, which minimise scorer variability and support stable measurement (Alderson, 2000; Hughes, 2003). Difficulty is controlled through text and task design rather than subjective judgement.

In speaking and writing, analytic rating scales with clearly defined, behaviourally anchored descriptors support consistent scoring by both human raters and automated systems. The use of common tasks across all candidates further reduces variability associated with task differences and enhances comparability.

Profile-based decision rules contribute to reliability by requiring consistency across multiple criteria, ensuring that overall level assignment reflects stable performance rather than isolated strengths or weaknesses.

8.4. AI-mediated scoring considerations

The design of the NEAS Test of English incorporates AI-mediated delivery and scoring, with corresponding implications for validity and reliability.

Scoring criteria are defined in terms of observable features of performance that can be consistently identified across both human and automated scoring contexts. These include features such as response development, lexical and grammatical control, fluency patterns, and discourse organisation. This ensures that automated scoring remains aligned with the construct rather than relying on indirect or opaque proxies (Xi, 2010; Chapelle & Chung, 2010).

At the same time, features are interpreted in relation to communicative effectiveness. Errors, hesitation, and variation in performance are evaluated in terms of their impact on meaning, preventing reduction of the construct to surface-level metrics.

Fairness in AI-mediated scoring is addressed through robustness to variation in accent, speech patterns, and linguistic background, supported by an intelligibility-based approach to pronunciation and careful control of task content.

Finally, the use of analytic criteria supports transparency and interpretability, allowing scores to be linked to identifiable aspects of performance rather than generated through opaque holistic models. This aligns with current best practice in technology-mediated assessment.

8.5. Summary of validity and reliability framework

The validity and reliability of the NEAS Test of English are supported by an integrated set of design principles. The test represents language ability through communicative performance across skills, minimises construct-irrelevant variance through controlled content and scoring, ensures reliability through objective and analytic scoring methods, and incorporates AI-compatible features without compromising construct integrity.

Together, these elements support a coherent interpretive argument in which test scores can be understood as meaningful, consistent, and defensible indicators of communicative language proficiency across CEFR levels.

9. Conclusion

The NEAS Test of English has been developed as a coherent, construct-driven assessment of language proficiency, grounded in contemporary theory and aligned with established international frameworks. Across all components of the test, design decisions have been guided by a consistent set of principles: that language ability is best understood as communicative performance; that proficiency is demonstrated through the integration of linguistic resources in meaningful tasks; and that assessment must balance validity, reliability, and practicality in real testing contexts.

The adoption of a four-skills framework ensures comprehensive coverage of language use, while the embedding of grammatical, lexical, discourse, and pragmatic competence within task performance reflects current understanding of how language operates in practice. The alignment with the CEFR provides an internationally interpretable basis for reporting outcomes, supported by clearly defined analytic criteria and transparent decision rules. Together, these elements ensure that reported scores represent meaningful differences in communicative ability rather than artefacts of test format or scoring procedures.

Particular attention has been given to maintaining construct validity while accommodating the constraints of AI-mediated, unsupervised delivery. The use of standardised tasks, observable scoring features, and analytic descriptors allows for consistent and scalable assessment without reducing the construct to superficial measures. At the same time, careful control of topic content, avoidance of accent bias, and the treatment of content as task responsiveness rather than topic knowledge contribute to fairness across diverse candidate populations.

The test's scoring and reporting framework further supports interpretability and defensibility. By combining analytic scoring with profile-based decision rules, the test avoids over-reliance on single dimensions of performance and ensures that overall proficiency reflects balanced ability across key aspects of language use. The inclusion of plus-level distinctions enhances diagnostic precision while remaining consistent with CEFR principles.

Taken together, the NEAS Test of English represents a principled implementation of communicative, CEFR-aligned language assessment in a contemporary delivery context. Its design reflects both theoretical coherence and practical applicability, ensuring that it can be used with confidence in educational and professional settings. By making its underlying assumptions, design choices, and interpretive framework explicit, the test supports transparency, accountability, and ongoing refinement in response to evidence and use.

10. Reference list

Alderson, J. C. (2000). Assessing reading. Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. Cambridge University Press.

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice. Oxford University Press.

Buck, G. (2001). Assessing listening. Cambridge University Press.

Cambridge English. (2015a). Cambridge English Scale: Research and validation. Cambridge Assessment.

Cambridge English. (2015b). Cambridge English Scale: Transforming the way English is assessed. Cambridge University Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1(1), 1–47.

Chang, A. C.-S. (2013). Effects of listening support on second language listening comprehension. TESOL Quarterly, 47(4), 863–883. https://doi.org/10.1002/tesq.97

Chang, A. C.-S., & Read, J. (2006).
The effects of listening support on the listening performance of EFL learners. TESOL Quarterly, 40(2), 375–397.

Chapelle, C. A., & Chung, Y. (2010). The promise of NLP and speech processing technologies in language assessment. Language Testing, 27(3), 301–315.

Cizek, G. J., & Bunch, M. B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Sage.

Council of Europe. (2001). Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.

Council of Europe. (2020). Common European framework of reference for languages: Companion volume. Council of Europe Publishing.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. TESOL Quarterly, 42(3), 475–493.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. TESOL Quarterly, 39(3), 379–397.

Ellis, R. (1994). The study of second language acquisition. Oxford University Press.

Field, J. (2008). Listening in the language classroom. Cambridge University Press.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), Examining listening: Research and practice in assessing second language listening (pp. 65–111). Cambridge University Press.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. College Composition and Communication, 32(4), 365–387.

Fulcher, G. (2003). Testing second language speaking. Pearson Longman.

Fulcher, G. (2010). Practical language testing. Hodder Education.

Fulcher, G., & Davidson, F. (2007). Language testing and assessment. Routledge.

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? Applied Linguistics, 35(5), 553–574.

Grabe, W. (2009). Reading in a second language: Moving from theory to practice. Cambridge University Press.

Green, A. (2014). Exploring language assessment and testing: Language in action. Routledge.

Harding, L. (2012). Accent and listening assessment: A review of the literature. Language Testing, 29(2), 163–190. https://doi.org/10.1177/0265532211415708

Harding, L. (2013). Accent and listening assessment. Language Testing, 30(2), 163–180.

Hughes, A. (2003). Testing for language teachers (2nd ed.). Cambridge University Press.

Hyland, K. (2004). Disciplinary discourses: Social interactions in academic writing. University of Michigan Press.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. Journal of Writing Research, 1(1), 1–26.

Khalifa, H., & Weir, C. (2009). Examining reading: Research and practice in assessing second language reading. Cambridge University Press.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in second language development. Applied Linguistics, 27(4), 590–619.

Levelt, W. J. M. (1989). Speaking: From intention to articulation. MIT Press.

McNamara, T. (1996). Measuring second language performance. Longman.

McNamara, T., & Roever, C. (2006). Language testing: The social dimension. Blackwell.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). Macmillan.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility. Language Learning, 45(1), 73–97.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? In P. Nation & M. Snowling (Eds.), Vocabulary in language teaching (pp. 59–82). Cambridge University Press.

O'Sullivan, B. (2011). Language testing: Theories and practices. Palgrave Macmillan.

Purpura, J. E. (2004). Assessing grammar. Cambridge University Press.

Read, J. (2000). Assessing vocabulary. Cambridge University Press.

Rost, M. (2011). Teaching and researching listening (2nd ed.). Routledge.

Shaw, S. D., & Weir, C. J. (2007). Examining writing: Research and practice in assessing second language writing. Cambridge University Press.

Sherman, J. (1997). The effect of question preview in listening comprehension tests. Language Testing, 14(2), 185–213. https://doi.org/10.1177/026553229701400204

Taylor, L., & Jones, N. (2006). Cambridge ESOL examinations and the CEFR. Research Notes.

Vandergrift, L., & Goh, C. (2012). Teaching and learning second language listening. Routledge.

Weigle, S. C. (2002). Assessing writing. Cambridge University Press.

Weir, C. J. (2005). Language testing and validation: An evidence-based approach. Palgrave Macmillan.

Weir, C. J., Hawkey, R., Green, A., & Devi, S. (2013). Examining listening: Research and practice in assessing second language listening. Cambridge University Press.

Xi, X. (2010). Automated scoring and feedback systems for spoken English. Language Teaching, 43(3), 291–308.