Scoring Writing & Speaking on the Duolingo English Test

Duolingo Research Report DRR-26-09
September 22, 2026 (51 pages)
https://englishtest.duolingo.com/research

Nitin Madnani, Phoebe Mulcaire, and Danwei Cai

Abstract

This report describes how the writing and speaking responses on the Duolingo English Test are scored. Our automated scoring approach is construct-driven and anchored in human expertise: assessment researchers decompose the constructs into relevant subconstructs, applied linguists and natural language processing (NLP) experts select the measurable response properties that provide supporting evidence of these subconstructs, and machine learning (ML) and artificial intelligence (AI) experts design the models that combine those properties into a score such that each subconstruct's contribution to a score can be measured. This gives us control over what the scoring systems respond to and the ability to verify that the resulting scores reflect the intended construct. We document this approach in detail, including the full set of features used and the safeguards applied for responses produced in bad faith.

This is a living document whose purpose is to provide up-to-date information about the Duolingo English Test, and it will be updated to reflect any major changes to our automated scoring methodology.

Keywords

Duolingo English Test, automated scoring, writing assessment, speaking assessment, machine learning, responsible AI

Corresponding author:
Nitin Madnani
Duolingo, Inc. 5900 Penn Ave, Pittsburgh, PA 15206, USA
Email: nitin@duolingo.com

Contents

1 Background

The **Duolingo English Test** (DET) is a computerized-adaptive, high-stakes English language proficiency assessment with the primary purpose of informing admissions decisions at English-medium universities (Naismith et al., 2026). The DET assesses the four language skills of speaking, writing, reading, and listening and is aligned to the proficiency levels of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001, 2020).

The DET is a *digital-first* assessment, meaning that it is designed for a digital environment from the start. As such, the DET uses machine learning technology and a human-in-the-loop AI approach throughout all test development, test scoring, and security processes. In this paper, we focus on one of these processes: the automated scoring of writing and speaking responses. Specifically, we describe how our approach to automated scoring is construct-driven and anchored in human expertise, retaining human control over what the systems measure and the ability to verify that the resulting scores reflect the intended construct. This allows the DET to score open-ended responses at scale without sacrificing the rigor expected of a high-stakes assessment.

The **writing** construct as measured by the DET is defined as the capacity to produce clear, well-organized, and appropriately complex written English across a range of purposes and audiences (Goodwin et al., 2025). This includes staying on topic, developing and organizing ideas, and using a range of accurate grammar structures and vocabulary. The DET defines the **speaking** construct as the capacity to speak clearly, fluently, and appropriately in English for academic and everyday situations (Park et al., 2025). The measurement of this construct from spoken responses focuses on their fluency, range of accurate grammar structures and vocabulary, logical structure of ideas, and intelligibility rather than on having a native-like or standard English accent. Each construct is measured through four open-ended task types whose responses are produced under time limits and scored automatically.

The remainder of this paper is organized as follows. §2 briefly introduces a few key types of automated scoring systems and the considerations that distinguish them. §3 describes the DET's approach to automated scoring at a high level. §4 and §5 then detail how writing and speaking responses are scored respectively, covering the constructs measured, the tasks used, the features used, and validity evidence showing how well our automated scores agree with expert human ratings and with scores from IELTS and TOEFL. §6 describes the governance, fairness, and human-oversight practices that surround our scoring systems, and §7 concludes. Three appendices provide supporting detail on the human rating process (Appendix A), the full catalog of scoring features (Appendix B), and the safeguards designed to catch bad-faith responses (Appendix C). Throughout the paper, technical notes set off from the main text provide implementation details for readers with a technical background; they can be skipped without loss of continuity.

2 A Brief Introduction to Automated Scoring

Automated scoring refers to the use of computer systems to evaluate and assign scores to test-taker responses (Beigman Klebanov & Madnani, 2022). Such systems have a long history in educational assessment, beginning with straightforward rule-based systems and evolving alongside advances in computing. An automated scoring system takes a response as input—whether typed text or recorded speech—and produces a score as output. Automated scoring systems can vary considerably, depending on the underlying approach

used. In subsequent subsections, we provide a high-level description of the most commonly used approaches. In brief:

- **Rule-based systems** apply explicit rules written by human experts (§2.1)
- **Machine learning (ML) systems** learn from data how to combine measurable properties of a response that human experts have defined (§2.2)
- **Deep learning (DL) systems** learn their own internal representations directly from data, *without* human-defined properties (§2.3)
- **Hybrid systems** combine two or more of these approaches (§2.4)

Irrespective of the underlying approach, automated scoring systems generally share the following goal: to assign a response the score that a qualified human expert would have assigned it. Expert human ratings are, therefore, central to automated scoring, both as the target that such systems are built to reproduce *and* as the benchmark against which their performance is judged. How closely an automated system's scores agree with expert human ratings is consequently the most direct evidence of how well it works.

However, agreement with human ratings—or **human-machine agreement**—is *not* the same as **construct validity**, i.e., whether the system actually measures the intended construct well. A system can reproduce expert scores closely while doing so on the basis of properties that have little to do with language proficiency—raw response length, say, or superficial features that happen to correlate with proficiency in the responses that were used to build the system. Establishing that a system measures what it is *meant* to measure requires more than just high human-machine agreement; it requires:

1. **Control:** being able to constrain in advance which properties of a response the system should use in determining the score, how they should be measured, and how much influence any one of them can have on the score.
2. **Verifiability:** being able to determine which properties the system *actually uses* to produce a given score, whether it behaves equitably across groups of test takers, whether it generalizes to responses unlike those used to build it, and whether it can be gamed.

These two considerations determine whether scores are valid, which is what matters for a high-stakes assessment. A secondary consideration for automated scoring systems is whether the scores can be explained to non-technical audiences without *any* specialized analysis. Some system designs may provide this affordance as a side benefit.

2.1 Rule-based systems

The simplest automated scoring systems are **rule-based**: human experts define explicit rules (e.g., "deduct a point if the response contains fewer than 50 words") and the system applies them mechanically. These systems afford complete control and verifiability: every property the system uses, how it is measured, and how much it influences the score are specified by hand. Their limitation is coverage: any manually created set of rules can only cover a fraction of the criteria that characterize proficient written and spoken English.

2.2 Machine learning systems

Systems a step above rule-based systems are **feature-based machine learning (ML) systems** (henceforth **ML systems**). In this approach, experts first identify measurable, construct-relevant properties of a response (e.g.,

content, vocabulary diversity, grammatical accuracy, or discourse coherence) and encode them as numerical **features**. For example, a simple feature representing *vocabulary diversity* (i.e., how much variety a test taker demonstrates in their word choice) may be encoded as the ratio of the number of unique words in the response to its total length in words, such that a response repeating the same few words receives a lower value than one drawing on a wider range. The defined features are then extracted for a large set of test-taker responses that have been **holistically rated** by human experts, i.e., assigned a single score reflecting the quality of the response as a whole rather than separate scores for individual aspects. An ML algorithm is then used to learn how to weight and combine the features to predict the experts' holistic ratings on this data.*

In ML systems, features are defined and interpreted by human experts, and the system's role is solely to learn the optimal combination of features via a mathematical function. Since the features are selected and defined by hand, control over which properties are used and how they are measured remains with human experts; what is learned is how those properties are combined. The simplest such function is a weighted sum, analogous to an overall course grade being computed from its components, e.g., homework average counting for 20%, a midterm counting for 30%, and a final exam counting for 50%. In this case, the contribution of each feature (component) to the final score (course grade) is fixed and easy to describe: it is simply the feature's weight. More complex functions can capture subtler interactions among features and often agree more closely with human ratings, but the effect of a change in any one feature then depends on the values of the others, so its contribution must be determined empirically rather than read off in advance. In short, the complexity of the combination function determines how much control remains over the influence of any individual feature, and how readily that influence can be verified.

2.3 Deep learning systems

A third category of automated scoring systems uses **deep learning**, a family of machine learning methods based on "deep" neural networks, i.e., with many layers of computation. These include **large language models (LLMs)** such as those powering generative AI tools like ChatGPT (OpenAI, 2025a). Where ML systems use complex functions to combine *human-defined* features, deep learning systems go a step further: they operate *directly* on raw text or audio without requiring human-defined features at all, instead learning their own internal representations of what matters in a response. For example, an end-to-end deep-learning automated scoring system can be trained to simply take a written or spoken response as input and produce a holistic score as output.

This may seem attractive due to reduced human effort and increased flexibility, but it cedes both control and verifiability. There are no human-defined properties to constrain: the system determines *for itself* what aspects of the response to use, and nothing limits how much any of it influences the score. This does not necessarily mean that such a system measures the construct poorly; in fact, it may well capture aspects of proficiency that hand-designed features miss. The difficulty is that this cannot be easily verified. While there is active research on explaining the behavior of deep learning models (Zhao et al., 2024), it remains substantially harder than with feature-based systems to determine why a particular score was assigned, or to verify that the score reflects the intended construct rather than spurious patterns in the data.

---

* The design of such features draws on a substantial body of work in automated scoring and computational linguistics; see Shermis and Burstein (2013) and Beigman Klebanov and Madnani (2022) for overviews covering written responses, and Zechner and Evanini (2019) for spoken responses.

2.4 Hybrid systems

While there are several examples of automated scoring systems employing solely rules, ML methods, or DL technology, there also exist **hybrid systems** that can combine these approaches (for a related overview, see Williamson, 2026). For example, an assessment developer may combine a rule-based component with an ML scoring system as follows: predefined rules assign the lowest possible score to non-English responses, while responses that pass this check go on to be scored by the ML system.

A second kind of hybrid system might be one that uses deep learning technology such as LLMs to produce a targeted measurement of a *single*, well-specified dimension of a response (e.g., how coherent the response is, or how clearly it is pronounced), and then incorporates that measurement as an individual feature in an ML scoring system. Here, the deep learning technology is used differently than in the *end-to-end* system we described previously. It is not being asked to fully judge the response on its own; instead, human experts decide *what* it should measure—a single property such as coherence—and train and validate it against expert human ratings of precisely that dimension. The component's internal workings remain opaque, as in any deep learning system, but its role is much more targeted: it makes a *specific* (not holistic) judgment about *one* property of the response, and its output is just *one* of many features the ML system combines to produce a score. This hybrid approach lets assessment developers take advantage of the strengths of deep learning technology while retaining control and verifiability: human experts decide what the deep learning component measures, its output is validated against human ratings of precisely that property, and its influence on the score is bounded by the ML framework used.

2.5 Summary

The four approaches to automated scoring described in the previous subsections differ in how much of the scoring process is specified by human experts and how much is learned from data. Moving from rule-based to end-to-end deep learning systems, progressively more is learned. This tends to increase agreement with human ratings, since the system can use whatever information in a response turns out to be predictive. *For the same reason*, it also reduces control and verifiability. Hybrid systems do not occupy a single point along this continuum: they combine components from the different approaches, and the amount of control retained depends on which components are learned and how the components are combined. Figure 1 summarizes the approaches and their ordering.

3 The DET Approach to Automated Scoring

The automated systems employed to score DET written and spoken responses—the *Duo Writing Scorer* and the *Duo Speaking Scorer*, respectively—are hybrid scoring systems, combining both types described in §2.4 and shown in Figure 1. They use ML models, with deep learning technology contributing some of the features and a rule-based component applied on top of the ML scores to handle bad-faith responses (see §4.2 and Appendix C). We provide additional high-level details about the systems here and delve deeper into them in subsequent sections. The process of creating and using DET automated scoring models is as follows:

1. **Defining subconstructs** — Our approach to automated scoring begins with construct definitions grounded in language assessment theory (Goodwin et al., 2025; Park et al., 2025). We then break down the speaking and writing constructs into relevant **subconstructs**, i.e., distinct, theoretically grounded components of the construct that capture a specific dimension of speaking or writing proficiency. For

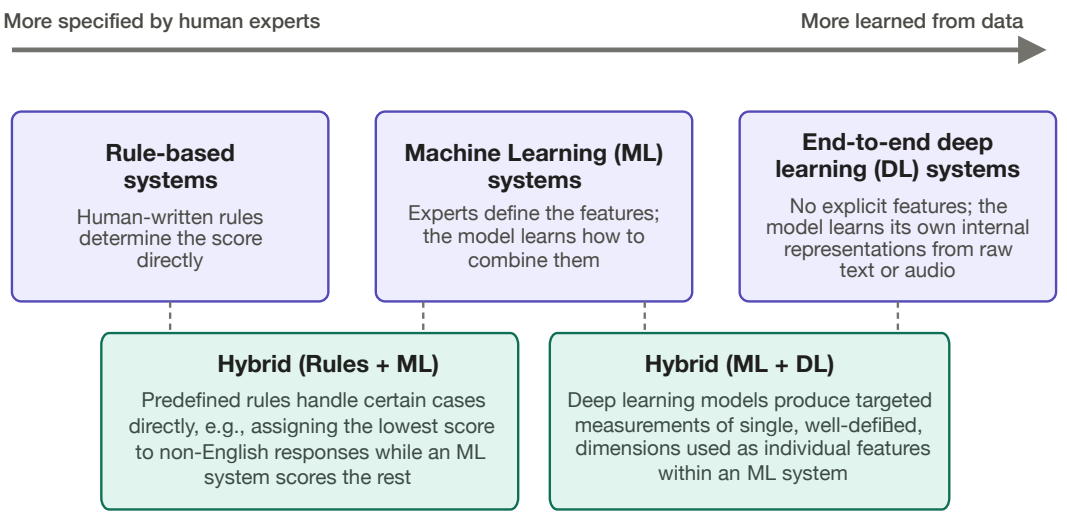

Figure 1. Three commonly used types of automated scoring systems (top) and two common ways they can be combined (bottom), ordered by how much of the scoring process is learned from data rather than specified by human experts. §2.5 describes what this ordering implies for human-machine agreement and for control and verifiability.

writing, we use four subconstructs: **content**, **discourse coherence**, **lexis**, and **grammar**. For speaking, we use the same four along with **fluency** and **pronunciation** as two additional subconstructs. These subconstructs are consistent with the dimensions assessed by other major automated scoring systems for writing (Quinlan et al., 2009; Lottridge et al., 2020, p. 6) and speaking (L. Chen et al., 2018; Zechner & Evanini, 2019), which converge on a similar set despite differences in terminology and grouping.

2. **Selecting & computing features** — For each subconstruct, experts select a set of measurable properties that, according to prior research, provide evidence of proficiency in that subconstruct. We then write computer programs to compute a numerical **feature** representing each such property. The choice of which features operationalize a subconstruct is itself a human decision, grounded in theory: a feature is included because it has a clear rationale linking it to the dimension of proficiency the subconstruct represents. For example, a potential feature for the content subconstruct might be the number of unique words in the response that are not shared with the prompt, a simplistic measure of how much *new content* the response contains, indicative of how well the response develops its ideas. Some features are computed using deep learning technology such as LLMs, but with a targeted scope as described in §2.4. Appendix B describes the full set of features used by the DET automated scoring systems, along with the relevant research citations.

3. **Training models** — Once we have programmed all of the features, we compute the entire set of features for a set of writing or speaking responses that have been scored by human experts (**training data**). These responses are drawn from a representative sample of the DET test-taker population and rated by qualified experts under ongoing quality control (§6 and Appendix A describe the composition of the

training data and human rating process in more detail). Next, we use ML techniques to build a **model** that produces automated scores aligned to the human experts' scores. In essence, this **trained model** is a mathematical function that defines how the features should be weighted and combined so as to align as closely as possible with the human expert scores, which serve as the primary learning target.

4. **Evaluating models** — The model is then evaluated to ensure the quality and fairness of the scores it produces. This includes:

   (a) examining whether the features associated with each subconstruct relate to each other as expected,
   (b) evaluating the trained model's predictive accuracy on human-rated response data that was not in the training data (i.e., held out) to confirm that the model can generalize beyond the specific responses it learned from,
   (c) examining correlations of the model's scores with external measures of English proficiency, such as IELTS and TOEFL scores, and
   (d) analyzing subgroup differences to verify that construct-irrelevant factors (e.g., gender, first language, or operating system) do not have any significant impact on the produced scores.

5. **Finalizing & deploying models** — At this point, the model is "frozen," meaning the feature computations and the mathematical function for combining them, learned during the training phase, are both fixed. This frozen model is then deployed for operational use: new writing and speaking responses submitted by test takers are analyzed to extract the same set of features as in step 3, which are then combined by the deployed model to produce the score. Any further changes to the features or the model can *only* happen through additional rounds of the human-governed training and evaluation process described above. In practice, this happens periodically as we identify parts of subconstructs that are under-covered and develop new features (or improve existing ones) to measure them better.

Figure 2 provides a summary of the steps in the DET automated scoring model development and deployment process as described above. Note that every component used in this process is defined, reviewed, and validated by human experts before it is ever deployed on the test. This includes the constructs being measured, the features that operationalize them, and the model that combines them. This is part of the DET's **human-in-the-loop AI** approach to automated scoring (Mosqueira-Rey et al., 2023; Munro, 2021); see Figure 22 of the DET technical manual (Naismith et al., 2026) for a visual overview of this approach.

It is equally important to be clear about what the DET does *not* do. Our scoring systems do *not* use deep learning technology, LLMs, or generative AI tools to make holistic judgments about *overall response quality.* We do not, for example, prompt ChatGPT to "rate this essay" and treat its output as the score. While we do make *targeted* use of some deep learning/LLM-based features *within* our ML scoring system (see §2.4 and Appendix B), every such feature is trained on and validated against human expert ratings of a specific, well-defined subconstruct. Additionally, such features are only a few among many others that feed into the scoring system. Every feature used in our models has a clear theoretical motivation tied to the construct being measured, and the behavior of the overall model can be examined and audited. This construct-driven, human-governed design is a deliberate choice: it ensures that scores are grounded in the principles of language assessment rather than the often unverifiable behavior of LLMs.

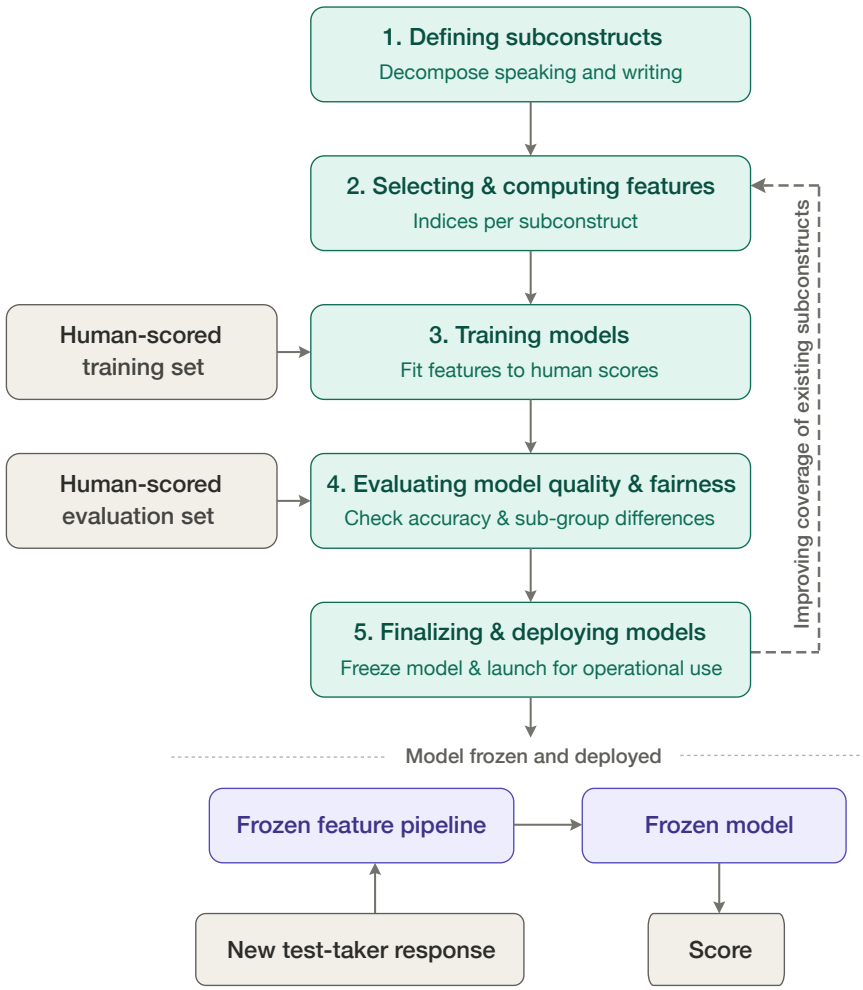

Figure 2. A high-level depiction of the general DET automated scoring model development and deployment process.

4 Scoring Writing Responses

The DET defines writing ability as the capacity to produce clear, well-organized, and appropriately complex written English across a range of purposes and audiences. This definition is grounded in the CEFR, with written production ranging from simple, formulaic texts at the A1 level to sophisticated, well-structured prose at the C2 level. The DET's writing construct also draws on sociocognitive frameworks (Mislevy, 2018; Weir, 2005)

and research on second language writing development (Goodwin et al., 2025). These frameworks recognize that performance on a writing task reflects not just writing ability but also related skills such as reading and pragmatic competence, general capacities such as critical thinking and content knowledge, and factors such as a test taker's confidence or familiarity with the test format.

To assess this construct, the DET includes four writing task types: *Interactive Writing*, *Write About the Photo*, *Interactive Listening (Summarization)*, and *Writing Sample*, described in Table 1. Each task is designed to elicit written responses that provide evidence of proficiency across different communicative contexts and purposes.

Table 1. Writing task types on the DET.

Task Type	*What test takers do*	*Time limit*
Interactive Writing	Test takers respond to a written prompt asking them to describe something, recount an experience, or argue a point of view. They then receive a follow-up prompt targeting themes their initial response did not fully address, encouraging further elaboration.	5 minutes (initial response) + 3 minutes (follow-up)
Write About the Photo	Test takers write a description of an image. Images depict people, animals, and objects in a wide range of contexts and are designed to give test takers the opportunity to use varied vocabulary and grammatical structures.	60 seconds
Interactive Listening (Summarization)	Following a multi-turn conversation set in a university context, test takers compose a written summary of the conversation.	75 seconds
Writing Sample	Test takers write an extended response to a prompt asking them to describe something, recount an experience, or argue a point of view. The response is shared with institutions alongside the test score.	5 minutes

4.1 Writing Subconstructs

The DET's automated scoring approach to writing operationalizes the construct through four subconstructs, each of which comprises finer-grained dimensions:

1. **Content**: The response is relevant to the task and develops its ideas substantively. Dimensions include task achievement, relevance, effect on the reader, appropriateness of style, and development.

2. **Discourse coherence**: The response is clearly structured and its ideas progress logically. Dimensions include clarity, cohesion, structure, progression of ideas, and appropriateness of format.

3. **Lexis**: The response draws on a range of vocabulary and uses it accurately and appropriately. Dimensions include lexical diversity, lexical sophistication, word choice, word formation, and error severity.

4. **Grammar**: The response uses a range of grammatical structures accurately and appropriately. Dimensions include range of grammatical structures, grammatical complexity, spelling, error frequency, error severity, and appropriateness.

These subconstructs are defined by task-specific holistic rubrics, which are aligned with CEFR descriptors, informed by relevant research (Goodwin et al., 2025), and reflect common elements shared with other established language proficiency assessments. The rubrics specify what characterizes performance at each proficiency level for each subconstruct, and serve as the basis for both the human ratings used to train the models and for the features designed to capture each dimension of writing quality. Appendix A provides a more comprehensive description of the human rating process.

The four subconstructs are operationalized through 90 carefully chosen numerical features drawn from the field of automated writing evaluation (AWE) using methods based on natural language processing (NLP) and computational linguistics (CL). Before features are extracted, responses are first **preprocessed**. At a high level, this includes identifying constituent parts (individual sentences and words) and labeling the identified words with their grammatical roles (verb, noun, adjective, etc.).

Technical Note. The preprocessing applied before feature extraction specifically includes sentence segmentation, tokenization (Jurafsky & Martin, 2009), and annotation with metadata such as dependency-tree relationships, lemmas (i.e., dictionary forms of words), and parts-of-speech (de Marneffe et al., 2021). Multiple tokenization schemes are used depending on the feature: some features operate over whitespace-delimited tokens, while others use more linguistically informed tokenization, e.g., from spaCy (Honnibal et al., 2020).

The features themselves span all four subconstructs. A few illustrative features for each subconstruct are listed below and shown in Figure 3, but a full list of all features with technical details and relevant citations is provided in Appendix B.

Content features

- *Prompt relevance:* Measures how closely the meaning of the response aligns with the meaning of the prompt.
- *Unique content length:* Counts the number of words and phrases the test taker produced beyond simply repeating words or phrases from the prompt.
- *Reference similarity:* For the *Interactive Listening (Summarization)* and *Write About the Photo* tasks, compares the response against high-quality reference summaries or captions to check whether key ideas are present.

Discourse coherence features

- *Sentence overlap:* Measures how much vocabulary is shared between consecutive sentences, which is one indicator of whether ideas flow logically from one to the next.
- *Semantic similarity between sentences:* Assesses whether adjacent sentences are related in meaning, not just in word choice.
- *LLM-based coherence rating:* Provides a holistic judgment of how well-organized and clearly structured the response is, as assessed by an LLM fine-tuned on human ratings of coherence.

Lexis features

- *Vocabulary diversity:* Measures the range of different words used in the response, distinguishing genuinely varied word choice from simple repetition.
- *CEFR-level word proportions:* Calculates the proportion of words at each CEFR proficiency level, capturing whether the test taker draws on more advanced vocabulary.
- *Mean word length:* Measures the average length of words in the response, since longer words tend to be rarer and more advanced.

Grammar features

- *Sentence complexity:* Measures the length and depth of grammatical structure within sentences, as an indicator of how varied and sophisticated the test taker's syntax is.
- *Grammatical error features:* A set of features based on detecting and classifying more than 50 error types (e.g., subject-verb agreement, article usage, and verb tense) using a proprietary error correction system. Each error type is a separate feature.
- *Differential use:* Captures patterns of word and grammar usage that statistically distinguish higher-proficiency writers from lower-proficiency writers, based on a large set of historical test-taker responses.

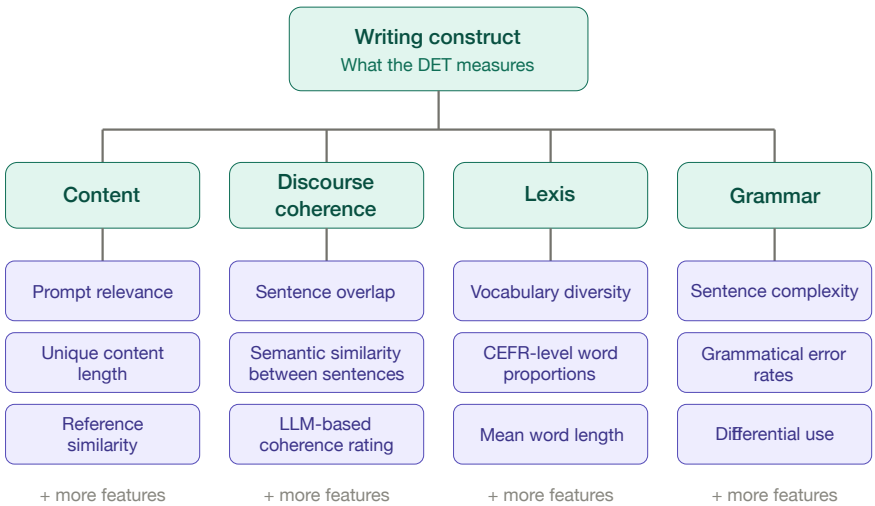

Figure 3. The DET writing construct, decomposed into four subconstructs, collectively operationalized through 90 carefully chosen numerical features. Three illustrative features are shown per subconstruct. See Appendix B for the full set of features.

4.2 Automated Scoring for Writing Tasks

To automatically score writing responses, the *Duo Writing Scorer* uses a separate ML model for each writing task type since different tasks were designed to elicit different kinds of evidence about writing ability (e.g., picture descriptions versus dialogue summaries). For each task type, we follow the procedure outlined in §3 to train, evaluate, and deploy the scoring model. These task-specific scoring models share the same overall design and draw on the same underlying feature set; what differs is that each model is trained on responses to its own task type, and that a few features apply only to certain task types. For example, comparison against reference summaries is relevant only for the *Interactive Listening (Summarization)* task type.

Technical Note. Each task-specific scoring model employs a two-level subconstruct-based architecture, where the two levels refer to the successive stages of ML models that are trained: one per subconstruct, and one that combines them.

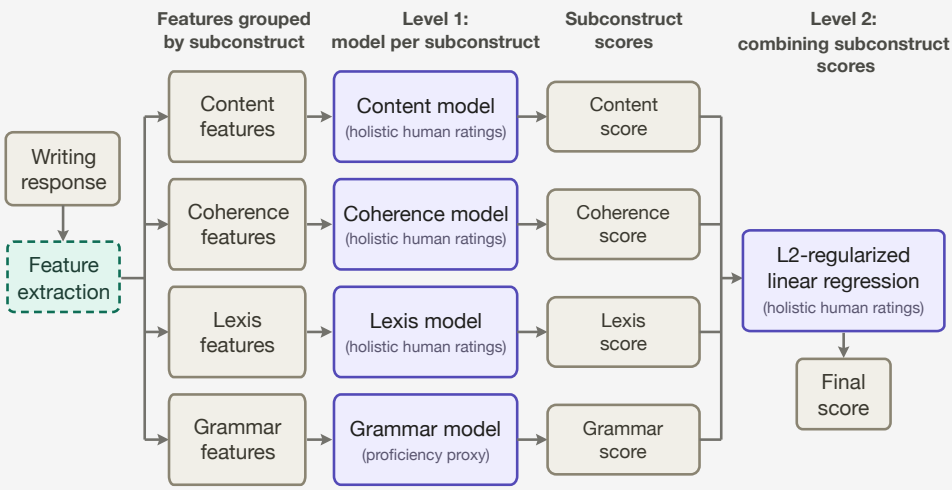

Figure 4. High-level diagram illustrating the two-level architecture of the models used to score written responses.

In the first level, features are grouped by subconstruct (content, discourse coherence, lexis, and grammar) and a separate ML model is trained for each, using only the features assigned to that subconstruct to produce a subconstruct score. The specific model class used at this level may vary across subconstructs and across grader versions (e.g., linear regression, XGBoost; T. Chen and Guestrin, 2016). The content, discourse coherence, and lexis subconstruct models are trained on holistic expert human ratings. Grammar is handled differently: since the grammar subconstruct comprises more than 50 individual error-type features, it requires more training data than the available human ratings can provide. We therefore train the grammar subconstruct model on a **proficiency proxy**—an estimate of the test taker's proficiency derived from their performance on the receptive (reading and listening) tasks on the test—which can be computed for a much larger pool of responses. In the second level, the

subconstruct scores are combined into a response-level score using an L2-regularized linear regression, also trained to predict holistic human ratings. This makes each subconstruct's contribution to the score directly inspectable.

This architecture provides several advantages:

1. *Verifiability.* Since the final combination step is a linear regression over subconstruct scores, the weight given to each subconstruct is directly visible through the regression coefficients, which are frozen once the model is trained. When these coefficients are multiplied by a response's subconstruct scores, it becomes straightforward to examine the contribution of various subconstructs to a particular response's score. A useful side benefit is that these contributions are also explainable: one can say plainly how much of a response's score came from grammar as opposed to content, without recourse to any specialized analysis.
2. *Control over the contribution of subconstructs.* The L2 regularization at the second level penalizes large coefficients, preventing the model from overly relying on any single subconstruct. This matters for construct validity: *all four* subconstructs contribute to the task score, rather than whichever one happens to be most predictive in the training data. Across writing task types, every subconstruct makes a meaningful contribution to the task score, and none dominates.
3. *Robustness.* The architecture provides a natural form of dimensionality reduction: rather than fitting a single model over the full feature space, the final regression operates over only the subconstruct scores, reducing the risk of overfitting to the responses in the training data.
4. *Control over the influence of deep learning technology.* Since features are partitioned by subconstruct, a feature derived from a deep learning technology such as an LLM can influence the final score only *through* the single subconstruct it is assigned to. Within that subconstruct, it competes with the features alongside it, *and* its total possible influence on the final score is capped by that subconstruct's second-level coefficient. This means that even a highly predictive LLM signal remains one piece of evidence among many rather than becoming a de facto holistic scorer.

Once the trained models have been deployed, they provide scores for the responses submitted for all writing tasks on the DET. However, some written responses are produced in **bad faith** rather than as genuine attempts at the task: they may be off-topic, written in a language other than English, unintelligible or incoherent, largely devoid of substantive content, or composed almost entirely of generic filler phrases. Since such responses are rarely seen in the training data and may still exhibit some surface characteristics associated with good-faith writing—e.g., reasonable length, varied vocabulary, and well-formed sentences—the ML models on their own may not assign them a sufficiently low score. Therefore, we apply a set of penalties on top of the model scores, each of which caps the score a response can receive. Some penalties are applied automatically, while others are raised as signals for the human proctor reviewing the test session, who confirms or rejects them before any adjustment is made. Appendix C describes these penalties in more detail.

A weighted average of the test taker's response scores across writing task types produces a **composite writing score**, using the weights shown in Table 2.

Table 2. Weights used to calculate the composite writing score. Each writing response receives its own score, and the composite is the weighted average of those scores.

Task Type	*Number of responses*	*Total weight*
Interactive Writing (Initial)	1	1/9
Interactive Writing (Follow Up)	1	1/9
Write About the Photo	3	3/9 (1/9 per response)
Interactive Listening (Summarization)	2	2/9 (1/9 per response)
Writing Sample	1	2/9

The composite writing score is then transformed into the final **DET Writing score** through a procedure called **scaling**. The scaling procedure maps the composite score—which is on the standard normal scale—to DET's 10–160 reported scale (in 5-point increments) using statistics derived from a large, representative sample of test-taker performance and ensures that the DET Writing score is comparable across test forms and over time (Nydick & Lockwood, 2025).

4.3 Validity Evidence

The DET Writing score is validated against multiple reference measures:

1. Human ratings produced by the DET's expert human raters and assigned on a six-point scale corresponding to A1–C2 on the CEFR (see Appendix A). Agreement with human ratings is the most direct test of whether the automated scoring captures the same construct that expert raters do.

2. Writing scores from two external English proficiency assessments, IELTS and TOEFL. Agreement with IELTS and TOEFL writing scores provides external validation.

Table 3 shows these agreement numbers for the DET Writing score, computed via Pearson's correlation.

Table 3. Pearson correlations between the DET Writing score and three reference measures: human ratings produced by DET expert raters, and writing scores from IELTS and TOEFL. Sample sizes (n) are shown in parentheses.

	Human Ratings	*IELTS Writing*	*TOEFL Writing*
DET Writing	.96 (n=350)	.60 (n=1540)	.72 (n=295)

Correlation values measure the strength and direction of a linear relationship between two measures and can range from -1 to 1; negative values indicate that as one measure increases, the other one tends to decrease, 0 indicates no relationship, and positive values indicate that as one measure increases, so does the other. The DET Writing score's correlation of .96 with human ratings, therefore, indicates very strong agreement with expert judgment. Correlations with IELTS and TOEFL are lower than those with human ratings, as expected: these are different tests, measuring overlapping but not identical constructs, and using different task types and rubrics.

A useful reference point is how closely IELTS and TOEFL writing scores agree with each other—reported at .68 (n=937) (Ikeda et al., 2025). The DET Writing score's correlations with both tests fall around that figure.*

Technical Note. The human ratings in Table 3 come from a full-session dataset comprising 2,800 expert ratings: for each of 350 test sessions, all eight writing responses were rated by a single expert rater. This makes it possible to compute a session-level human writing score using the same task weights applied to the automated scores (Table 2). Because both measures aggregate over eight responses, each is more reliable than a score for any single response, which is why this correlation is higher than the task-level correlations reported later in Table 4.

Since the DET Writing score is a composite over multiple open-ended task types, it is also worth examining how well the automated system's scores agree with human ratings at the task level. Table 4 shows Pearson correlations computed over individual responses, grouped by task type, between automated writing scores and human ratings (**human–machine**), alongside the corresponding correlations between two human raters on the subset of responses that were double-rated (**human–human**).† The latter is useful since it serves as a good benchmark for automated scoring performance: if an automated system agrees with an expert rater as well as a second expert does, then its judgments are consistently aligned with human judgments of proficiency. Across the five scored writing response types, the human–machine agreement is close to or above the corresponding human–human benchmark. Together with the score-level results in Table 3, this indicates that the automated scores align with expert judgment not only in aggregate but also at the level of individual task types.

Table 4. Pearson correlations between the DET task-level automated writing scores and human ratings (human–machine) and between two human ratings (human–human). Sample sizes (n) are shown in parentheses.

Task Type	*Human–machine correlation* (r)	*Human–human correlation* (r)
Interactive Writing (Initial)	.90 (n=3150)	.89 (n=890)
Interactive Writing (Follow Up)	.88 (n=3150)	.89 (n=630)
Write About the Photo	.83 (n=3641)	.85 (n=1217)
Interactive Listening (Summarization)	.89 (n=2954)	.90 (n=700)
Writing Sample	.87 (n=3398)	.82 (n=2307)

5 Scoring Speaking Responses

The DET defines speaking ability as the capacity to speak clearly, fluently, and appropriately in English for academic and everyday situations. As with writing, this definition is grounded in the CEFR, with spoken production ranging from simple, formulaic utterances at the A1 level to sophisticated, well-structured extended discourse at the C2 level. The DET's speaking construct draws on the same sociocognitive frameworks as writing, and additionally incorporates research on second language speech production, since speaking

* Note that the DET Writing score is on a different scale (10–160) than the human ratings or the IELTS and TOEFL scores, so the correlation values reflect how similarly the measures order test takers rather than how closely the numbers themselves match.
† Double-rating is performed on a random subset of responses as part of the rating quality control process; see Appendix A.

involves real-time processes such as articulation, pacing, and self-monitoring that do not apply to written responses (Park et al., 2025).

To assess the speaking construct, the DET includes four different speaking task types. Three of these (*Speak About the Photo*, *Read, then Speak*, and *Speaking Sample*) are monologic and require the test taker to provide a single sustained spoken response. The fourth, *Interactive Speaking*, is dialogic: it requires the test taker to participate in a simulated communicative exchange with an on-screen avatar across two topics. Because the tasks differ in their structure, they use different scoring pipelines: the three monologic task types share a single scoring approach (see §5.2), while *Interactive Speaking* is scored through a related but distinct pipeline designed to handle its adaptive administration (see §5.3). Table 5 summarizes the four task types.

Table 5. Speaking task types on the DET.

Task Type	*What test takers do*	*Time limit*
Speak About the Photo	Test takers give a spoken description of an image. Images depict people, animals, and objects in a wide range of contexts.	90 seconds
Read, then Speak	Test takers read a written prompt and give an extended spoken response describing something, recounting an experience, or arguing a point of view.	90 seconds
Speaking Sample	Test takers give an extended spoken response to a prompt asking them to describe something, recount an experience, or argue a point of view. The recording is shared with institutions alongside the test score.	3 minutes
Interactive Speaking	Test takers participate in a multi-turn spoken conversation with an on-screen avatar across two topics. Each topic begins with an initial question, followed by two to three follow-up questions selected adaptively based on real-time evaluation of the test taker's earlier responses. Each session contains six to eight questions in total.	35 seconds per response

5.1 Speaking Subconstructs

The DET's automated scoring approach to speaking operationalizes the construct through six subconstructs. Four of these (**content**, **discourse coherence**, **lexis**, and **grammar**) are shared with writing (see §4 for details). The two additional subconstructs are specific to speaking:

1. **Fluency**: The test taker speaks smoothly and at a natural pace. Dimensions include the pace of speech, the grouping of words into phrases, and overall smoothness of delivery.
2. **Pronunciation**: The test taker's speech can be easily understood in terms of its *intelligibility* (how accurately a listener recognizes the words) and *comprehensibility* (how much effort understanding requires). Dimensions include individual sounds, word and sentence stress, rhythm, intonation, and connected speech. Note that this subconstruct does *not* assess how closely a test taker approximates any particular variety or accent of English; clearly intelligible speech is scored as such regardless of the accent.

These subconstructs are defined by task-specific holistic rubrics which are aligned with CEFR descriptors, informed by relevant research (Park et al., 2025), and reflect common elements shared with other established language proficiency assessments. The rubrics serve as the basis for both the human ratings used to train the models and the features designed to capture each dimension of speaking proficiency. Appendix A provides a more comprehensive description of the human rating process.

The six subconstructs are operationalized through 120 carefully chosen numerical features drawn from the fields of automated writing evaluation (AWE) and automated speech evaluation (ASE) using methods based on natural language processing (NLP) and computational linguistics (CL). Because speaking responses are produced as audio rather than text, preprocessing begins with an additional step: each response is first transcribed to text using **automatic speech recognition (ASR)**, which produces a verbatim transcript along with the timing of each spoken word and a measure of how confident the system was in transcribing it.

We use verbatim transcripts because they preserve features of natural, unrehearsed speech that are necessary for measuring fluency, such as filler words ("um," "uh"), repeated words, and moments where the test taker restarts or corrects themselves. The presence of such phenomena correlates with speaking proficiency—in fact, they are *exactly* what the fluency features are designed to capture—but they arise from the cognitive demands of producing speech in real time rather than from a test taker's lexical or grammatical ability. Therefore, the fluency features are computed from the verbatim transcript, while the content, discourse coherence, lexis, and grammar features are computed from a "clean" version of the transcript. This version of the transcript has disfluencies removed and is preprocessed the same way as written responses (identifying sentences, words, and grammatical roles). This sharing of features across the writing and speaking modalities is intended: the same features can be used to predict expert human judgments of qualities such as coherence and grammatical accuracy in both written and spoken responses, and using them across both ensures that these qualities are measured consistently.

Technical Note. The ASR model used to score the monologic speaking tasks is a fine-tuned version of NVIDIA's Parakeet model (Sekoyan et al., 2025), trained on a corpus of L2 English speech to produce verbatim transcriptions of test-taker responses. Beyond the verbatim transcript, the ASR model also produces (a) word- and phoneme-level timing information used by the fluency features, and (b) intermediate acoustic representations (from the model's acoustic encoder) used as input to compute the holistic pronunciation score feature (see Appendix B). To produce the cleaned transcript used for content, discourse coherence, lexis, and grammar feature extraction, disfluencies are identified using a fine-tuned mmBERT model (Marone et al., 2025) that assigns each token a probability of being a disfluency, in conjunction with a predefined filler-word list. Importantly, this cleaning preserves the test taker's original L2 lexical and grammatical forms; it removes filler words, repetitions, and self-repairs but does not normalize or correct the underlying language.

In addition to the features shared with writing, the *Duo Speaking Scorer* includes a set of speaking-specific features designed to operationalize the fluency and pronunciation subconstructs. A few illustrative features for these speaking-specific subconstructs are listed below and shown in Figure 5, but a full list of all features with technical details and relevant citations is provided in Appendix B. Features for content, discourse coherence, lexis, and grammar are computed from the transcript in the same way as for writing and are not repeated here.

Note that even though these features are shared between the *Duo Writing Scorer* and the *Duo Speaking Scorer*, the actual ML models trained using the features are entirely separate, so the same feature may carry a different weight in the speaking model than in the writing model.

Fluency features

- *Speech rate:* Measures how quickly the test taker speaks, in words or syllables per second, after accounting for pauses.
- *Pause patterns:* Captures the frequency, duration, and location of silences within the response, distinguishing natural phrasing pauses from hesitations that disrupt the flow of speech.
- *Fluency score:* Provides a holistic judgment of fluency from a deep learning model trained to match human ratings of fluency, taking the full response into account.

Pronunciation features

- *Speech recognition confidence:* Uses the ASR model's own confidence in its transcription as an indicator of how intelligible the speech is; less intelligible speech produces lower-confidence transcripts.
- *Pronunciation score:* A holistic score from a deep learning model trained to match human ratings of how clearly and intelligibly the test taker produces English sounds, word stress, and sentence intonation.
- *Prosodic features:* A set of features capturing how the voice rises, falls, and varies in emphasis across speech, capturing the stress and intonation patterns that contribute to intelligible speech.

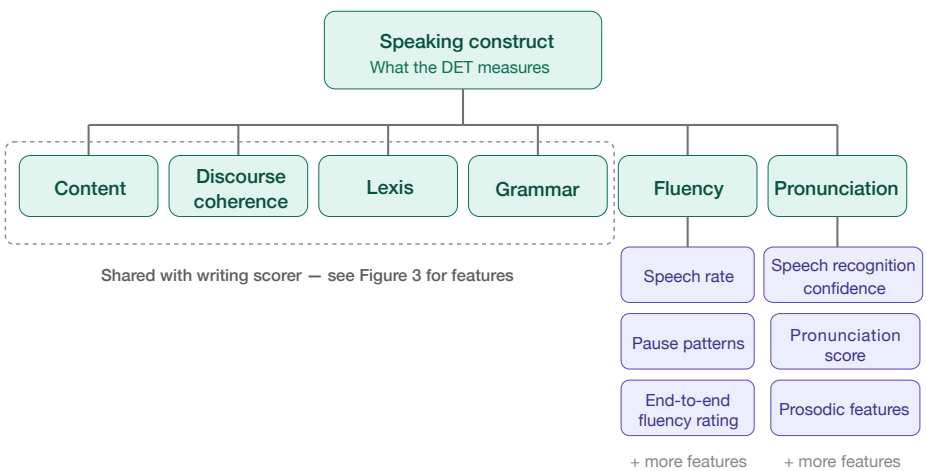

Figure 5. The DET speaking construct, decomposed into six subconstructs (four shared with writing, two speaking-specific), collectively operationalized through 120 carefully chosen features. Three features per subconstruct are shown for illustration. See Appendix B for the full set of features used.

5.2 Automated Scoring for Monologic Speaking Tasks

The approach the *Duo Speaking Scorer* uses to automatically score responses to the three monologic speaking tasks is similar to the one used for the writing tasks. Features are first extracted on a large training set of spoken responses that have been scored independently by expert human raters using the CEFR-aligned rubrics. The extracted features are then combined into task-specific ML models. Once trained, the models are evaluated for quality and fairness and, assuming no issues are discovered, frozen for deployment. This evaluation includes the predictive accuracy and fairness checks described for writing models in the previous sections.

Technical Note. The ML models used to score the monologic speaking tasks employ the same two-level subconstruct-based architecture described in §4.2, extended from four subconstructs to six: content, discourse coherence, lexis, and grammar are joined by fluency and pronunciation.

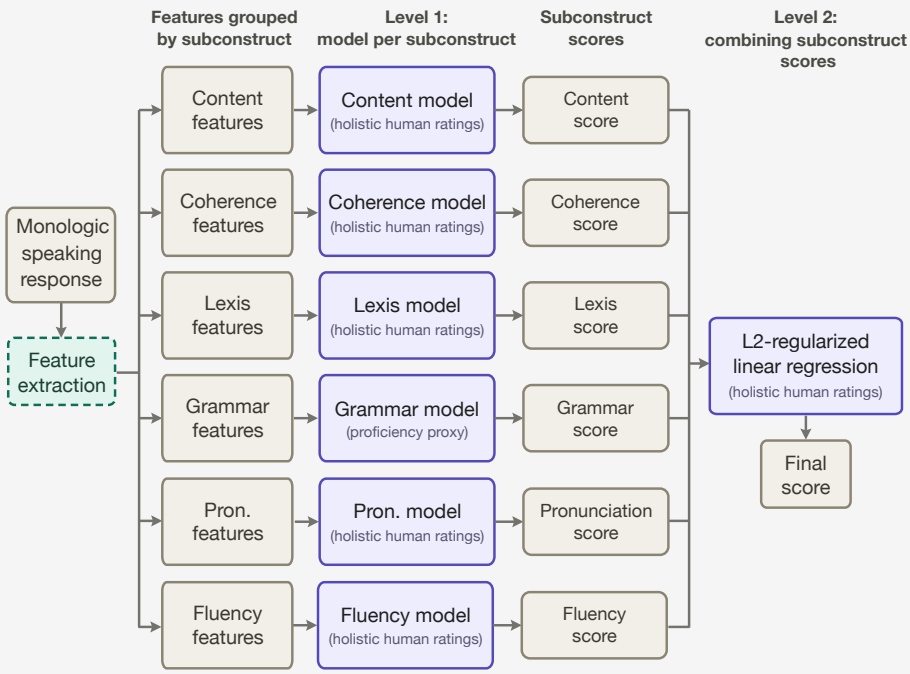

Figure 6. High-level diagram illustrating the two-level architecture of the ML models used to score monologic spoken responses.

As with writing, the specific model class used at the first level may vary across subconstructs, while the second level is always a linear model. The learning targets are also the same as for writing: each subconstruct model is trained on holistic expert human ratings, except grammar, which is trained on a proficiency proxy for the reason described in §4.2. For the monologic speaking models, the proxy is derived from the test taker's performance on the test's reading, listening, and writing tasks.

The second level likewise combines the six subconstruct scores into a response-level score using an L2-regularized linear regression trained to predict holistic expert human ratings.

The advantages of this architecture are as described in the technical note in §4.2. Two are particularly relevant here:

- *Control over the contribution of subconstructs.* Pronunciation features are highly predictive of human ratings and could otherwise dominate the score at the expense of other subconstructs; the L2 regularization at the second level constrains this. Across speaking tasks, every subconstruct makes a meaningful contribution to the task score, and no single subconstruct dominates the others.
- *Control over the influence of deep learning technology.* The *Duo Speaking Scorer* relies more heavily on deep learning technology than the *Duo Writing Scorer* since it includes end-to-end ratings of fluency and pronunciation as features (see Appendix B). Therefore, bounding these features' influence within their respective subconstructs is more consequential here.

5.3 Automated Scoring for Interactive Speaking Task

Interactive Speaking on the DET is a novel speaking task that differs from the three monologic speaking tasks in two important ways. First, it is *dialogic*: the test taker participates in a multi-turn conversation rather than producing a single sustained response. Second, it is *adaptively administered*: within each topic, follow-up questions are selected from a predefined bank based on a real-time evaluation of the test taker's earlier responses on that topic (Park et al., 2026). These differences motivate an automated scoring system that is different but still closely related to the one described in §5.2.

An Interactive Speaking task consists of two topics. A topic has one initial question and two to three follow-up questions, for a total of six to eight questions in a test session. All topics, questions, and the question-specific rubrics described below are generated in advance using an LLM and comprehensively reviewed by human experts before deployment on the test.

During the test, a test taker's response to each question is evaluated in real time against a question-specific checklist consisting of two to five key points: one to two *critical* key points required for a successful response, and one to three *important elaborations* that demonstrate broader task completion. The proportion of key points covered by the response, as judged by an LLM, is computed as the task completion score for that response. As an illustrative example, for a question like *"Tell me about a hobby you enjoy,"* a critical key point might be "names a specific hobby" and an elaboration might be "explains why they enjoy it." The task completion score is used in two ways:

1. **Adaptive administration**: the system maintains a running estimate of the test taker's task completion scores across their prior responses and uses it to select the next follow-up question from one of three difficulty buckets (easy, medium, hard), while also avoiding follow-ups whose content the test taker has already covered.
2. **Scoring**: task completion is used to define the content subconstruct when assigning scores to the test-taker responses (described next).

After the test session is complete, each Interactive Speaking response is scored along the same six subconstructs used for the monologic speaking tasks—content, discourse coherence, lexis, grammar, fluency, and pronunciation—with one major difference: the content subconstruct for Interactive Speaking is realized *primarily* through the task completion score, with only one additional response-level relevance feature. The remaining five subconstructs are operationalized similarly to monologic speaking (see §5.2 and Appendix B).

As with the monologic tasks, these features are then fed into an ML scoring system. Since Interactive Speaking is adaptively administered, not all test takers answer questions of equivalent difficulty. Therefore, raw scores are not directly comparable across test takers: the same score indicates greater ability on a harder question than on an easier one. To account for this, we take the difficulty of each question into account when estimating the overall score for each subconstruct. These six subconstruct scores are then weighted and combined into a single score for the Interactive Speaking task.

Technical Note. The Interactive Speaking scoring pipeline generally follows the same two-level subconstruct-based architecture as the monologic models (§5.2), with two notable differences: (1) an additional adjustment is applied to the scores to account for adaptive administration, and (2) the learning targets used are different.

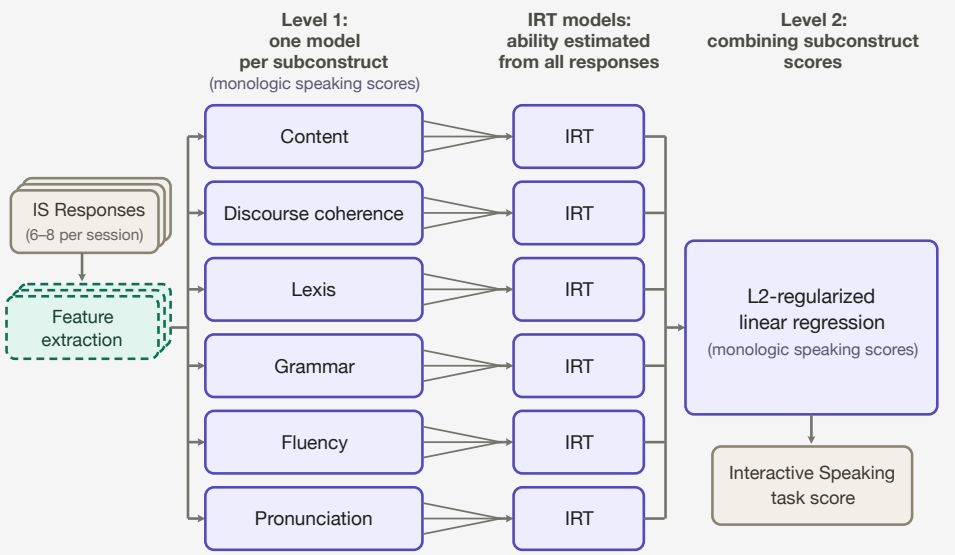

Figure 7. High-level diagram illustrating the two-level architecture of the ML models used to score the Interactive Speaking task.

In the first level, a separate ML model is trained for each subconstruct to produce a raw per-response subconstruct score from the relevant features. As with the monologic models, the specific model class may vary across subconstructs. What differs here is the learning target: rather than holistic expert

human ratings, every subconstruct model is trained to predict an average of the automated scores produced by the monologic speaking models.

In order to make scores comparable across test takers who answered questions of varying difficulty, we use an **Item Response Theory (IRT)** model for continuous responses. For each subconstruct, the scores from all six to eight responses in the session are modeled together, with each response treated as a separate item whose difficulty has been estimated in advance, to produce a single estimate of the test taker's ability on that subconstruct. This yields six IRT-based subconstruct scores per test session, rather than per response.

In the second level, the six IRT-based subconstruct scores are weighted and combined into the Interactive Speaking task score. The weights are informed by fitting an L2-regularized linear regression that predicts the average monologic speaking score from the six subconstruct estimates on held-out data.

Despite these differences, the overall design preserves the verifiability of the two-level architecture; each subconstruct's contribution to the task score can be explicitly measured while still accommodating the adaptive administration.

5.4 Proctor Review & Composite Speaking Scores

Unlike writing responses, spoken responses are not subject to automated scoring penalties; instead, all DET spoken responses are reviewed by human proctors as part of the DET's test security process. Proctors may flag responses that indicate a test taker did not attempt the speaking tasks in good faith: producing no speech at all, reading from a prepared script, or giving a response unrelated to the question asked.

The threshold for applying these flags is deliberately high. For unrelated responses in particular, proctors are instructed to:

- target *only* deliberate attempts to circumvent our automated scoring, e.g., delivering a memorized answer *regardless* of what was asked,
- *not* apply the flag to lower-scoring test takers, since the purpose of the speaking section is to assess speaking ability rather than the ability to follow instructions,
- err on the side of leniency whenever *any* connection can be found between the prompt and the response, and
- *not* flag responses where a test taker has simply misheard or misunderstood the prompt.

When a proctor does flag a response, the session is invalidated and no score is reported. A weighted average of the response scores across the test taker's speaking tasks produces a **composite speaking score**, using the weights shown in Table 6.

Similar to the writing case, the composite speaking score—which is on the standard normal scale—is transformed into the final **DET Speaking score** through a procedure called **scaling**, which maps the composite score onto the 10–160 scale.

Table 6. Weights used to calculate the composite speaking score. Each monologic speaking response receives its own score, while Interactive Speaking yields a single score for the whole task (see §5.3); the composite is the weighted average of these scores.

Task Type	*Number of scores*	*Total weight*
Speak About the Photo	1	1/5
Read, then Speak	1	1/5
Speaking Sample	1	1/5
Interactive Speaking	1	2/5

5.5 Validity Evidence

We present the validity evidence for the automated DET Speaking score in this section, complementing §4.3 which presented similar evidence for writing. Table 7 shows the agreement with human ratings (on a six-point scale) and with IELTS and TOEFL speaking scores for the DET Speaking scores. The comparison with human ratings is restricted to the monologic speaking tasks, since human ratings are not available for Interactive Speaking over the same set of test sessions, which makes it difficult to compute a session-level human rating covering all four tasks. The comparisons with IELTS and TOEFL use the full DET Speaking score, including Interactive Speaking.

Unlike the writing comparison in §4.3, the human ratings in Table 7 do not come from a single, full-session dataset. Responses to each of the three monologic speaking tasks were rated separately, and the ratings were averaged over the test sessions for which at least two were available. Because this averages over multiple responses, it is more reliable than a score for any single response, which is why this correlation is higher than the task-level correlations reported in Table 8.

Table 7. Pearson correlations between the DET Speaking score and three reference measures: human ratings produced by DET expert raters (denoted by an asterisk as it excludes Interactive Speaking), and speaking scores from IELTS and TOEFL. Sample sizes (n) are shown in parentheses.

	*Human Ratings**	*IELTS Speaking*	*TOEFL Speaking*
DET Speaking	.92 (n=3534)	.66 (n=1099)	.68 (n=444)

Recall from §4.3 that correlation values measure the strength and direction of a linear relationship between two measures and can range from -1 to 1.

The DET Speaking score's correlation of .92 with human ratings indicates very strong agreement with expert judgment. Correlations with IELTS and TOEFL are lower, potentially due to differences in the constructs, task types, and rubrics. Ikeda et al. (2025) report that IELTS and TOEFL speaking scores correlate with each other at .69 (n=937). The DET Speaking score's correlations with both tests fall around that figure.*

*As with writing, the measures being compared are on different scales, so the values reflect how similarly they order test takers rather than how closely the numbers match.

Like the DET Writing score, the DET Speaking score is also a composite over multiple open-ended task types. Therefore, we also present evidence of how well the automated scores for speaking agree with human ratings at the task level. Table 8 shows Pearson correlations computed over individual responses, grouped by task type, between automated speaking scores and human ratings (**human–machine**) alongside the corresponding correlations between two human raters on the subset of responses that were double-rated (**human–human**).

Table 8. Pearson correlations between the DET task-level automated speaking scores and human ratings (human–machine) and between two human ratings (human–human). Sample sizes (n) are shown in parentheses.

Task Type	*Human–machine correlation* (r)	*Human–human correlation* (r)
Speak About the Photo	.87 (n=2709)	.88 (n=757)
Read, then Speak	.87 (n=2642)	.85 (n=742)
Speaking Sample	.86 (n=2637)	.85 (n=792)
Interactive Speaking	.84 (n=4008)	.90 (n=600)

Across the three monologic tasks, the human–machine correlations fall within a narrow band, indicating that automated scoring performs consistently across task types. Comparing them to human–human correlations shows that the automated scores agree with a human expert about as closely as a second human expert does. The correlation between Interactive Speaking scores and human ratings is somewhat lower—likely because its subconstruct models are trained to predict the output of the other speaking models rather than human ratings directly—but remains high in absolute terms.

6 Governance

The DET's scoring systems are automated but not autonomous. Every score the test reports is the output of a measurement pipeline that human experts designed, trained, validated, and continue to monitor.

Automated scoring is one part of a larger AI ecosystem on the DET, which also includes automated item (or question) generation, item calibration, and remote proctoring. The DET's Responsible AI (RAI) standards (Burstein, 2026a) articulate the ethical principles that govern AI use across that ecosystem: Validity & Reliability, Fairness, Privacy & Security, and Accountability & Transparency. Each proposed change to a scoring system must be documented, evaluated against the RAI standards through an internal review process before deployment, and retained as a record that can be examined later. The DET's Responsible AI transparency report (Burstein, 2026b) documents how these standards are applied across the ecosystem.

The rest of this section describes the practices that implement these principles for automated scoring: §6.1 describes the human oversight that underpins the human-in-the-loop development and mandatory empirical validation; §6.2 discusses the control and verifiability that support accountability and transparency; and §6.3 describes the demographic representation and subgroup analyses required under fairness. Privacy and security are addressed through separate measures across the test as a whole, described in the DET security report (Belzak et al., 2025).

6.1 Human oversight

Each part of the writing and speaking scoring pipelines described in the previous sections rests on human decisions. Specifically:

- **Construct definition.** What it means to write or speak proficiently in English on the DET is articulated by applied linguists and assessment researchers, grounded in the CEFR and contemporary L2 development theory, and documented in publicly available construct whitepapers (Goodwin et al., 2025; Park et al., 2025). The subconstructs—content, discourse coherence, lexis, grammar, fluency, and pronunciation—and the rubrics that operationalize them are products of multiple years of assessment research expertise and experience.
- **Feature engineering.** The features that are used in the scoring systems are selected by computational linguists and AI scientists for their alignment with the construct. Each feature is documented and has a clear link to one of the six subconstructs. Appendix B lists the features individually, with their rationale, subconstruct association, and technical implementation.
- **Model training and validation.** The training data for the ML models is assembled from a representative sample of test takers (see §6.3) and rated by qualified human experts whose work is itself subject to quality control (see Appendix A). Models are trained on this data, and every new model is empirically evaluated in multiple ways before being deployed: (a) for predictive accuracy against held-out human ratings, (b) for agreement with external proficiency measures, and (c) for fairness across demographic subgroups.
- **Continuous monitoring.** Once deployed, DET scoring systems are continuously monitored. Score distributions, behavior on edge cases, and subgroup performance are tracked over time, and systems are retrained or updated when we observe drift or when the test itself undergoes changes (new questions, changes in population, etc.).

The takeaway is that no DET score is produced by a system that operates *outside* human-defined boundaries. The construct, the features, the training signals for the ML models, the validation criteria, and the penalty thresholds are *all* set by humans.

6.2 Control and verifiability

In §2, we described how automated scoring systems differ in the control and verifiability they afford, depending on how much of the scoring process is specified by human experts. Hybrid systems span a range in this respect, and the DET's scoring systems are deliberately constructed to sit at the more controlled end of it: hand-crafted, construct-aligned features feed into ML systems whose contributions can be quantified at both the feature and subconstruct level, and the deep learning components that do appear are confined to producing individual features.

Technical Note. Referring to the typology of automated scoring systems detailed in Williamson (2026), the DET's scoring systems are most closely aligned with feature-based engines trained on human-rated responses (Route B in their framework), with the qualification that a small number of features

are themselves the outputs of fine-tuned deep-learning models or LLMs (e.g., the GPT-based coherence & content rating features as well as the end-to-end pronunciation and fluency scores). However, the overall architecture remains verifiable with every feature named, documented, and clearly mapped to a subconstruct.

Control is exercised at two points. First, features are selected in advance by human experts, so the system can only respond to those aspects of the response that experts chose to measure. Second, the chosen architecture limits the influence any one feature or subconstruct can have on the score (see the technical note in §4.2).

Verifiability matters for two reasons. First, when needed, the scoring systems can be interrogated. The features that contributed to a given score, and their relative weights at the subconstruct level, can be easily inspected. For more complex feature combinations *within* a subconstruct, established techniques for explaining model predictions can be used to determine the contributions of each individual feature.* Second, for fairness analysis, it is important to know which features have the biggest impact on scores to assess whether any of those features behave differently across test-taker groups.

6.3 Fairness

Fairness on the DET is ensured not by a single check but rather by a set of relevant practices:

- **Representation in training data.** The test-taker samples used to train the ML scoring systems are taken from the operational DET population and reviewed for balanced representation across demographic variables such as first language (L1) and gender to mitigate bias. The main training samples comprise roughly 100,000 test sessions each for speaking and writing. L1s are quite diverse; test takers report over 100 first languages, and the most common first language accounts for under 17% of sessions. Gender representation in every dataset is close to evenly balanced in terms of male and female, with a small proportion specifying that their gender falls outside of the provided options.
- **Subgroup performance evaluation.** New ML scoring models are evaluated for differential performance across demographic subgroups defined by variables including L1, gender, and geographic region. We examine whether test takers of comparable proficiency receive systematically different scores depending on subgroup membership. We also compute agreement between automated and human scores within each subgroup, look for systematic over- or under-prediction that might indicate the model is using construct-irrelevant signals, and compare the score distributions the new and current models produce for each subgroup. Models are also evaluated at the transitions between adjacent CEFR levels, where linguistic differences (e.g., fluency, syntactic complexity, lexical range) are most nuanced and reliable discrimination can be hardest.
- **Feature-level checks.** Newly introduced features undergo differential feature functioning analysis before being incorporated into a model. This analysis examines whether a feature behaves differently for test takers of comparable proficiency depending on their subgroup membership. For example, the end-to-end pronunciation score was examined for bias across gender, operating system, and language family

* A widely used example is Shapley values (Lundberg & Lee, 2017), which assign each feature a contribution score for a given prediction.

groups. These investigations inform feature engineering decisions and occasionally lead to features either being updated or deprecated.

The analyses described here pertain to the scoring models; the DET also conducts differential item functioning analyses on test items themselves, which are reported in the DET Technical Manual (Naismith et al., 2026). The published research describing individual scoring features documents the feature-level analyses in more detail; see, for example, Cai et al. (2025) for the pronunciation score and Naismith et al. (under review) for the fluency features.

7 Conclusion

The DET's automated scoring of writing and speaking responses is not a set of algorithms replacing human judgment, but rather systems designed to scale expert judgment to match the needs of a digital-first assessment.

Every part of the pipeline that produces the DET Writing and Speaking scores is the collaborative effort of applied linguists, language assessment experts, psychometricians, computational linguists, and AI engineers. This includes:

- the construct definitions grounded in CEFR and language assessment theory,
- the rubrics that operationalize each subconstruct,
- the features designed to capture the dimensions those rubrics describe,
- the human ratings that train the ML models,
- the validation criteria the ML models must meet before deployment, and
- the monitoring and fairness checks that continue after deployment.

Automation enters this pipeline only *after* these parts of the human-defined framework are in place, and in a manner such that its contributions can always be inspected and audited.

This design reflects a deliberate choice. As discussed in §2, automated scoring systems differ in how much of the scoring process is specified by human experts and how much is learned from data, and this determines how much control and verifiability they provide. Systems that use LLMs or other deep learning models to make *holistic* judgments of proficiency cede both: there are no human-defined features to constrain and it is not easy to establish exactly what contributed to a given score. The DET's approach to automated scoring combines human expertise and machine learning to retain control and verifiability while still optimizing human-machine agreement. Features are aligned to specific subconstructs. The models that combine them are structured so that each subconstruct's contribution to a score can be measured. Deep learning technology is used only to produce a few specific, well-defined features.

For institutional stakeholders, the practical implication is that DET scores are produced with the same level of expertise that goes into producing scores on traditional, human-rated assessments, but applied at the scale a modern, digital-first assessment requires.

A DET Human Ratings

The human raters for the DET are experts in the field of English language teaching and assessment. All raters are required to have five or more years of TESOL experience with adults, two or more years of experience in speaking and writing assessment, TESOL certification, a relevant bachelor's degree or equivalent, expert English proficiency (C2 on the CEFR), and familiarity with the CEFR. Experience in high-stakes assessment, an advanced degree in a relevant field, and advanced certifications such as the Cambridge DELTA are preferred, and raters typically exceed the minimum requirements in most categories. Raters are also selected to ensure gender balance and a diversity of nationalities, geographic locations, linguistic backgrounds (including English varieties), and work experiences.

To rate writing and speaking responses, raters use task-specific holistic scale rubrics. These rubrics are publicly available and assess all relevant subconstructs. The scale has six points, corresponding to A1–C2 on the CEFR; the descriptors were developed by adapting the CEFR's own level descriptors to the specific demands of each task type, while also being informed by the rubrics used by other established assessments such as IELTS and TOEFL. Before being used for large-scale rating, each rubric is piloted using benchmark rating activities conducted by DET project leads, who apply the draft descriptors to sample responses and revise them where they prove difficult to apply consistently. Additional categories for bad-faith responses are also used. For datasets collected to develop features targeting specific subconstructs (e.g., content), trait-specific rubrics are used instead.

Before working on any rating project, raters attend training sessions for the task types they will rate—overall writing, overall speaking, or individual subconstructs—and must pass a certification test. Both initial certification and ongoing quality control are based on three metrics.

- **Exact agreement** is the proportion of responses on which two raters assigned identical scores.
- **Adjacent agreement** is the proportion on which their scores were either identical or differed by a single scale point.
- **Quadratic Weighted Kappa (QWK)** measures the agreement between two sets of ratings *after* accounting for the agreement that would be expected by chance alone. A QWK value of 0 indicates agreement no better than chance and 1 indicates perfect agreement. Unlike exact and adjacent agreement, it weights disagreements by their magnitude, so a disagreement of two points on the scale counts more heavily against it than a disagreement of one point.

We double-rate a minimum of 15% of all responses, which provides the paired ratings these metrics are computed from. Raters must meet target thresholds on these metrics to certify, and must continue to meet them to remain active.

Once certified, raters are regularly calibrated against **benchmark responses**: the responses carrying the gold-standard ratings produced during the rubric piloting described above, which serve as agreed-upon exemplars for each response type. A small calibration dataset is also rated by every active rater, which confirms continued scoring consistency and provides the linking data required for the statistical analyses described below.

Rater performance is also monitored statistically. Each year, we conduct Many Facet Rasch Measurement (MFRM) analyses to assess the consistency of human ratings and the performance of the rating scales, alongside a qualitative survey eliciting rater feedback on the rubrics and the rating process. The score

distribution data and category statistics from these analyses show that the speaking and writing scales are well-functioning and that raters are within acceptable ranges for both **severity** (how strictly a rater rates relative to the pool) and **reliability** (how consistently they apply the scale) for all tasks and skills. Results are shared with individual raters as feedback on how their rating tendencies compare with the pool and change over time.

B Writing & Speaking Features

This appendix provides a full description of the features used by the DET automated writing and speaking scoring systems. Each table shows the feature groups that contribute to a specific writing or speaking subconstruct—one or more features that measure the same underlying property in slightly different ways. Within each table, entries are sorted alphabetically by their group name. Each entry includes a high-level description accessible to a general audience, followed by technical implementation details for the interested reader.

Note that some speaking features whose implementation depends on the ASR system (e.g., the holistic pronunciation and fluency scores) are computed differently for the monologic and Interactive Speaking scoring pipelines.*

Content

Table 9. Content features used to score writing and speaking.

Feature group *list of features*	**Details**
Content Rating *gpt-content-rating*	**High-level Description:** Provides a holistic judgment of how well the response develops its content: whether it stays on topic, addresses the prompt fully, and elaborates ideas in a coherent and substantive way. It is produced by an LLM that has been specifically fine-tuned to match expert human ratings of content development. **Technical Implementation:** This feature uses a GPT-4.1-mini model (OpenAI, 2025b) fine-tuned to predict human ratings of content development. The prompt, response, and a short rubric with instructions are passed to the model, and a rating is extracted from the predicted completion. Different fine-tuned models are used depending on the task type: 1. For the *Interactive Writing* and *Writing Sample* tasks, we fine-tune one model using human ratings of content development obtained for responses to both tasks. 2. For the *Interactive Listening (Summarization)* task type, we use the same fine-tuned model as in (1), but with a different prompt. 3. For the *Write About the Photo* task, we fine-tune a model on human ratings of content development obtained for that task's responses, but with reference image captions provided as additional input. **Relevant Citations:** OpenAI (2025b)

* Interactive Speaking currently uses Whisper (Radford et al., 2023) for ASR, which provides different transcripts compared to the *verbatim* ASR used for the monologic tasks (§5.1). This means that while the feature *definitions* in this appendix are indeed shared across both scorers, some features—notably ones that rely more heavily on the internals of the ASR system, e.g., the holistic pronunciation and fluency scores—are computed differently for the two pipelines.

Table 9 continued

Feature group *list of features*	**Details**
Prompt Relevance *idf-embeddings-relevance*, *sentran-relevance*, *sentran-relevance-initial-prompt*, *sentran-relevance-response-context*	**High-level Description:** Measures how closely the meaning of the response aligns with the meaning of the prompt or topic. For the *Interactive Writing* task, the relevance of the follow-up response is measured against *both* the initial prompt and the initial response, since the follow-up is a continuation of the original exchange rather than a self-contained task. **Technical Implementation:** Two complementary approaches are used to measure the semantic similarity between a response and its prompt: • *Sentence-transformer embeddings.* For the *Writing Sample* and *Interactive Writing* tasks, the response and prompt are encoded using a pretrained neural sentence-transformer (Reimers & Gurevych, 2019) model (*paraphrase-MiniLM-L6-v2*), which encodes sentences so that similar meanings receive similar embeddings, and the feature *sentran-relevance* is computed as the cosine similarity between the two embeddings. For the *Interactive Writing* follow-up response, two additional features measure relevance to the other anchors of the exchange: *sentran-relevance-initial-prompt* is the cosine similarity between the follow-up response embedding and the embedding of the initial prompt, and *sentran-relevance-response-context* is the cosine similarity between the follow-up response embedding and the embedding of the initial response. • *IDF-weighted word embeddings.* For the *Read, then Speak*, *Speaking Sample*, and *Interactive Speaking* tasks, the response and prompt text are each represented as the IDF-weighted average of fastText word embeddings (Mikolov et al., 2017), with inverse document frequency (Sparck Jones, 1972) giving more weight to rare, topic-distinguishing words. For the *Speak About the Photo* task, the prompt embedding is derived from the mean of embeddings for 10 image captions generated by GPT-4 (Achiam et al., 2023). In all cases, the feature *idf-embeddings-relevance* is the cosine similarity between the response and prompt representations. This approach has been shown to measure topical relevance effectively in a learner-essay setting (Rei & Cummins, 2016), and we apply it here to transcribed spoken responses. **Relevant Citations:** Sparck Jones (1972); Rei and Cummins (2016); Mikolov et al. (2017); Reimers and Gurevych (2019)
Reference Similarity *semantic-similarity*	**High-level Description:** For the *Interactive Listening (Summarization)* and the *Speak/Write About the Photo* tasks, the response is compared to a set of high-quality reference responses to check whether the key ideas are present. **Technical Implementation:** For the *Interactive Listening (Summarization)* task, the feature is the maximum semantic similarity between embeddings of the response and each of ten reference responses to a question. Embeddings are produced by a pretrained neural embedding model (*paraphrase-MiniLM-L6-v2*), which encodes sentences so that similar meanings receive similar embeddings (Reimers & Gurevych, 2019). For the *Speak/Write About the Photo* tasks, GPT-generated captions are used in place of reference responses. **Relevant Citations:** Reimers and Gurevych (2019)

Table 9 continued

Feature group *list of features*	**Details**
Task Completion Score *gpt-task-completion*	**High-level Description:** For the *Interactive Speaking* task only, measures how comprehensively the response addresses the question against a question-specific rubric of *key points*. Each rubric defines two to five elements that a successful response should contain: one to two *critical* key points (required for a successful response) and one to three *important elaborations* (additional content that demonstrates broader task completion). The feature is the proportion of these key points present in the response, as judged by an LLM. This is the *primary* feature that defines the content subconstruct for Interactive Speaking. **Technical Implementation:** The rubric and response transcript are passed to a fine-tuned GPT-4.1-mini (OpenAI, 2025b) model that evaluates each key point as either present or absent in the response. The feature value is the number of key points judged present divided by the total number of key points in the rubric. The same score is also computed in real time during the test session and used to drive adaptive selection of follow-up questions (see §5.3). **Relevant Citations:** OpenAI (2025b)
Unique Content Length *num-characters*, *num-words*, *num-unique-2-grams*, *num-unique-3-grams*, *num-main-assertions*	**High-level Description:** Counts the amount of content the test taker produced, measured in characters and words, as well as in unique two- and three-word sequences that do not also appear in the prompt. The unique-sequence variants reward test takers who produce genuine new content while not rewarding those who write highly repetitive responses or copy from the prompt. **Technical Implementation:** *num-characters* and *num-words* are straightforward counts over the response. *num-unique-2-grams* and *num-unique-3-grams* count the unique bigrams and trigrams that appear in the response but not in the prompt, helping measure content development while limiting credit for repetition or prompt-copying. *num-main-assertions* is the count of main clauses in the response, identified using a dependency parse; this is more robust, especially for spoken responses since verbatim ASR transcripts may contain unreliable comma usage. **Relevant Citations:** Attali and Burstein (2006)

Discourse Coherence

Table 10. Discourse coherence features used to score writing and speaking.

Feature group *list of features*	**Details**
Coherence Rating *gpt-coherence*	**High-level Description:** Provides a holistic judgment of how well-organized and clearly structured the response is, produced by an LLM that has been specifically trained to match human ratings of coherence. **Technical Implementation:** This feature uses a GPT-4.1-mini model (OpenAI, 2025b) fine-tuned to predict human ratings of discourse coherence. The prompt, response, and a short rubric with instructions are passed to the model, and a 0–6 rating is extracted from the predicted completion. The method and rubric are similar to those described in Naismith et al. (2023), but further fine-tuning achieves higher agreement with human ratings, approaching inter-human agreement. The fine-tuning process follows a structure similar to that of the *gpt-content-rating* feature. **Relevant Citations:** Naismith et al. (2023); OpenAI (2025b)
Semantic Similarity Between Sentences *coh-lsa-local-similarity-mean*, *coh-lsa-local-similarity-stdev*, *coh-lsa-global-similarity-mean*, *coh-lsa-global-similarity-stdev*	**High-level Description:** Checks whether adjacent sentences are related in meaning, not just in the specific words they share. This captures coherence that comes from ideas flowing together conceptually, even when the sentences use different vocabulary. Both the average level of relatedness and the consistency of that relatedness across the response are measured. **Technical Implementation:** Pairwise semantic similarity between sentences is computed using Latent Semantic Analysis (LSA), following the approach of McNamara and Graesser (2012). The four features summarize the distribution of these pairwise similarities in two dimensions: scope and statistics. • *Local* features are computed over adjacent sentence pairs, capturing how smoothly the ideas flow from one sentence to the next. • *Global* features are computed over all sentence pairs in the response, capturing overall idea cohesion. • *Mean* features (*coh-lsa-local-similarity-mean*, *coh-lsa-global-similarity-mean*) report the average pairwise similarity, indicating the typical level of relatedness. • *Standard-deviation* features (*coh-lsa-local-similarity-stdev*, *coh-lsa-global-similarity-stdev*) report the variability in pairwise similarity, indicating whether relatedness is consistent across the response or fluctuates. **Relevant Citations:** McNamara and Graesser (2012)

Table 10 continued

Feature group *list of features*	**Details**
Sentence Overlap *coh-local-binary-noun-overlap*, *coh-local-binary-arg-overlap*, *coh-local-binary-stem-overlap*, *coh-local-proportion-content-overlap*, *coh-local-coreference-overlap-mean*, *coh-global-binary-noun-overlap*, *coh-global-binary-arg-overlap*, *coh-global-binary-stem-overlap*, *coh-global-proportion-content-overlap*, *coh-global-coreference-overlap-mean*	**High-level Description:** Measures how much vocabulary and referential content is shared between sentences in the response, indicating whether ideas flow logically from one sentence to the next. Responses in which adjacent sentences share related words, or refer back to the same entities, tend to be more coherent. **Technical Implementation:** A set of ten features that measure cohesion between pairs of sentences via two complementary mechanisms. *Lexical overlap.* Eight features measure cohesion by computing proportions of overlapping words between sentence pairs (McNamara & Graesser, 2012). Features vary along three dimensions and are named accordingly as *coh-[scope]-[computation]-[token_type]-overlap*: • **Scope**: *local* variants compute the mean overlap over consecutive sentence pairs; *global* variants compute the mean over all sentence pairs. • **Computation**: *binary* variants score each sentence pair as 1 if any overlap exists and 0 otherwise; *proportion* variants compute the number of unique tokens in each sentence that also occur in the other. • **Token type**: *noun* compares tokens tagged as nouns; *arg* compares tokens tagged as nouns or pronouns; *stem* compares lemmatized nouns in the second sentence against the set of lemmatized nouns, verbs, adverbs, and adjectives in the first; *content* compares only content words. Not all combinations are instantiated: the *binary* variants use *noun*, *arg*, and *stem*, while the *proportion* variants use textitcontent only, giving eight lexical-overlap features in total. *Coreference overlap.* Two additional features measure cohesion through coreference rather than direct lexical match: entity mentions (including pronouns and other referring expressions) are identified across the response using AllenNLP (Gardner et al., 2018; Lee et al., 2018) and grouped into coreference chains. The features compute the average overlap of these chains between pairs of sentences over adjacent sentence pairs (*coh-local-coreference-overlap-mean*) and over *all* sentence pairs (*coh-global-coreference-overlap-mean*). These features capture cohesion that arises when sentences refer to the same entity using different surface forms (e.g., an entity introduced as “the professor” and later referenced as “she”). **Relevant Citations:** McNamara and Graesser (2012);Gardner et al. (2018); Lee et al. (2018)

Fluency

Table 11. Fluency features used to score speaking.

Feature group *list of features*	**Details**
Fluency Score *e2e-sequence-fluency-score*	**High-level Description:** Provides a holistic judgment of fluency from a deep learning model trained to match human ratings of fluency, taking a structured symbolic sequence derived from the full spoken response as input. **Technical Implementation:** A deep neural network sequence model trained on construct-aligned human fluency ratings. The model takes as input a multi-stream token sequence consisting of the phoneme labels, the POS tags (plus special disfluency tags), and the phoneme/silence durations from forced alignment. The output is the feature value: a predicted fluency score aligned with human ratings. **Relevant Citations:** Naismith et al. (under review)
Pauses (between clauses) *inter-clause-silence-per-clause*, *inter-clause-silence-speech-ratio*, *inter-clause-silence-mean*, *inter-clause-silence-mad*, *inter-clause-0.8s-silence-per-clause*, *inter-clause-0.8s-silence-speech-ratio*, *inter-clause-0.8s-silence-mean*, *inter-clause-0.8s-silence-mad*, *inter-clause-0.8s-silence-ratio*, *inter-clause-0.8s-silence-duration-ratio*	**High-level Description:** Captures pauses at clause-like boundaries, including all such pauses as well as the *long* pauses (longer than 0.8 seconds). Pausing at clause boundaries is a natural feature of fluent speech, but excessively long or frequent pauses of this type can still signal difficulty with planning what to say next. **Technical Implementation:** A set of features derived from silences detected at clause-like boundaries in the verbatim ASR transcript, computed both over all such silences and over the subset longer than 0.8 seconds. *inter-clause-silence-per-clause* is the average number of silences per clause. *inter-clause-silence-speech-ratio* is the ratio of total inter-clause silence duration to speech duration. *inter-clause-silence-mean* is the mean duration of inter-clause silences. *inter-clause-silence-mad* is the mean absolute deviation of silence durations. The four corresponding *0.8s* variants apply the same calculations to the subset of pauses exceeding 0.8 seconds. *inter-clause-0.8s-silence-ratio* is the proportion of inter-clause silences that exceed the 0.8-second threshold, and *inter-clause-0.8s-silence-duration-ratio* is the proportion of total inter-clause silence duration accounted for by long silences. **Relevant Citations:** Yan et al. (2025)

Table 11 continued

Feature group *list of features*	**Details**
Pauses (between sounds) *inter-phoneme-silence-per-phoneme*, *inter-phoneme-silence-speech-ratio*, *inter-phoneme-silence-mean*, *inter-phoneme-silence-mad*, *inter-phoneme-0.8s-silence-per-phoneme*, *inter-phoneme-0.8s-silence-speech-ratio*, *inter-phoneme-0.8s-silence-mean*, *inter-phoneme-0.8s-silence-mad*, *inter-phoneme-0.8s-silence-ratio*, *inter-phoneme-0.8s-silence-duration-ratio*	**High-level Description:** Captures pauses between individual sounds in the response, including all such pauses as well as the *long* pauses (longer than 0.8 seconds) that are typically perceived as disfluent hesitations rather than natural breaks. Frequent or unusually long pauses between sounds tend to suggest that the speaker is having difficulty producing continuous speech. **Technical Implementation:** A set of features derived from silences detected between phonemes in the ASR output, computed both over all inter-phoneme silences and over the subset of silences longer than 0.8 seconds. *inter-phoneme-silence-per-phoneme* is the average number of silences per phoneme. *inter-phoneme-silence-speech-ratio* is the ratio of total inter-phoneme silence duration to speech duration. *inter-phoneme-silence-mean* is the mean duration of inter-phoneme silences. *inter-phoneme-silence-mad* is the mean absolute deviation of silence durations (i.e., the average distance of each silence from the mean silence duration), capturing variability in pause length. The four corresponding *0.8s* variants apply the same calculations to the subset of pauses exceeding 0.8 seconds. *inter-phoneme-0.8s-silence-ratio* is the proportion of inter-phoneme silences that exceed the 0.8-second threshold, and *inter-phoneme-0.8s-silence-duration-ratio* is the proportion of total inter-phoneme silence duration accounted for by long silences. **Relevant Citations:** Yan et al. (2025)
Pauses (within clauses) *within-clause-silence-count-mean*, *within-clause-silence-per-word-mean*, *within-clause-silence-speech-ratio-mean*	**High-level Description:** Captures pauses that occur in the middle of a clause-like segment, reflecting the kind of mid-phrase hesitation that often signals a speaker is searching for a word or formulating an idea. Pauses inside a clause typically interrupt the flow of speech more than pauses at clause boundaries. **Technical Implementation:** A set of features derived from silences detected within clause-like segments in the ASR transcript. *within-clause-silence-count-mean* is the average number of silences per clause. *within-clause-silence-per-word-mean* is the average number of within-clause silences per word. *within-clause-silence-speech-ratio-mean* is the average ratio of within-clause silence duration to speech duration across clauses. **Relevant Citations:** Yan et al. (2025)

Table 11 continued

Feature group *list of features*	**Details**
Speech Rate *effective-speech-duration*, *effective-speech-ratio*, *verbalize-words-per-second*, *syllables-per-second*, *mean-length-of-run*	**High-level Description:** Measures how quickly and continuously the test taker speaks. These features capture *speed fluency*—the pace at which a speaker produces language—and indicate whether the test taker can sustain the rate of speech expected at higher proficiency levels. **Technical Implementation:** A set of features derived from the ASR transcript and word- and phoneme-level timestamps. *effective-speech-duration* is the cumulative amount of time within the response during which speech is detected, after excluding non-speech intervals. *effective-speech-ratio* is the proportion of the total response duration spent in active speech. *verbalize-words-per-second* is the number of words spoken per second of total response duration. The word count excludes punctuation and tokens tagged as fillers or repairs. *syllables-per-second* is the analogous rate computed at the syllable level, providing a more granular measure of articulation rate that is less sensitive to differences in average word length. *mean-length-of-run* is the average number of syllables produced between pauses, capturing how much speech the test taker can produce in a single uninterrupted stretch. **Relevant Citations:** Yan et al. (2025)

Grammar

Table 12. Grammar features used to score writing and speaking.

Feature group *list of features*	**Details**
Differential Use of Grammar *du-pos*, *du-dep*	**High-level Description:** Captures patterns of grammar usage that statistically distinguish higher-proficiency writers from lower-proficiency writers. A response that uses grammatical structures in ways characteristic of higher-proficiency writing receives a higher value, while one whose grammar patterns resemble lower-proficiency writing receives a lower value. **Technical Implementation:** Based on n-gram classifiers trained on a large corpus of historical test-taker responses, split into high- and low-proficiency groups by median proficiency proxy. For each task type, one n-gram model is trained on high-proficiency responses and another on low-proficiency responses. The feature is computed as the log-odds ratio of the response being produced by a high-proficiency vs. low-proficiency test taker, based on the length-normalized probabilities from each model. Two variants are computed: *du-pos* (a 6-gram model of part-of-speech tags) and *du-dep* (a 6-gram model of dependency-tree tags). Models use interpolated modified Kneser-Ney smoothing via the KenLM library (Heafield et al., 2013) to deal with the small number of observations for longer n-grams. Mathematically, the log-odds ratio is defined as $V = \log[P(C{=}H \mid R)/P(C{=}L \mid R)] = \log[P_{\text{ngram-}H}(R)/P_{\text{ngram-}L}(R)]$, where R is the response, C is the proficiency class of the test taker who wrote the response, $P_{\text{ngram-}H}$ is the high-proficiency n-gram model's length-normalized probability of R, and $P_{\text{ngram-}L}$ is the low-proficiency n-gram model's length-normalized probability of R. We further regularize V to account for noise in the n-gram model's probabilities for very short responses, allowing the feature value to smoothly transition from a (slightly negative) default value for such responses to the expected log-odds ratio V as the responses get longer. The smoothed value S is defined as $S = V + (D - V) \times min(1.0, T/L)$, where D is the minimum default value, T is the minimum length threshold, and L is the response length. **Relevant Citations:** Attali (2011); Yancey et al. (2021); Vajjala and Lučić (2018); Heafield et al. (2013)

Table 12 continued

Feature group *list of features*	**Details**
Grammatical Error Rates *neg-[type]-errors-per-word* (**52 features**)	**High-level Description:** Detects grammatical errors in the response across 52 distinct types (e.g., subject-verb agreement, article usage, and verb tense) and measures how frequently each type of error occurs relative to the length of the response. **Technical Implementation:** The response is corrected using Duolingo's proprietary grammatical error correction model, based on mBART (Liu et al., 2020) and trained to correct common L2 learner mistakes. Each correction is classified into one of 52 non-spelling error types using the ERRor ANnotation Toolkit (ERRANT; Bryant et al., 2017). Error counts for each type are tabulated and divided by the total number of tokens in the response to produce an error-rate feature per type. **Relevant Citations:** Bryant et al. (2017); Liu et al. (2020)
Sentence Complexity *mean-sentence-length*, *max-dep-depth*, *mean-dep-depth*, *mean-main-assertion-length*, *mean-main-assertion-dep-depth*, *max-main-assertion-dep-depth*	**High-level Description:** Measures the length and depth of grammatical structure within sentences, as an indicator of the variety and sophistication of the English syntax used in the responses. Longer sentences and sentences with more deeply nested grammatical structures tend to reflect higher proficiency. **Technical Implementation:** A set of features computed as summary statistics over sentence-level grammatical structure, calculated in two ways. The first set of features uses sentence boundaries identified by punctuation: *mean-sentence-length* is the mean number of tokens per sentence, while *mean-dep-depth* and *max-dep-depth* are the mean and maximum dependency-tree depth across sentences, where dependency-tree depth is the length of the longest path from a leaf to the root token in the dependency parse of each sentence. The second set of features substitutes main clauses for sentences as the unit of analysis. Main clauses are identified from the dependency parse, making the resulting features more robust for spoken responses where verbatim ASR transcripts may contain unreliable comma usage. *mean-main-assertion-length* is the mean number of tokens per main clause, and *max-main-assertion-dep-depth* is the maximum dependency-tree depth across main clauses. *mean-main-assertion-dep-depth* is computed differently: rather than averaging per-clause maxima, it is the mean dependency depth across all tokens in the response. **Relevant Citations:** Schwarm and Ostendorf (2005)
Spelling Error Rate *neg-R:SPELL-errors-per-word*	**High-level Description:** Counts the number of spelling errors in the response relative to its total length. **Technical Implementation:** Spelling errors are identified using Duolingo's proprietary grammatical error correction model (based on mBART; Liu et al., 2020) and classified using ERRANT (Bryant et al., 2017). The feature is computed as the count of spelling errors divided by the total number of tokens in the response. **Relevant Citations:** Bryant et al. (2017); Liu et al. (2020)

Lexis

Table 13. Lexis features used to score writing and speaking.

Feature group *list of features*	**Details**
CEFR-Level Word Proportions *lex-comp-a2*, *lex-comp-b1*, *lex-comp-b2*, *lex-comp-c1*, *lex-comp-c2*, *lex-comp-oov*	**High-level Description:** Measures what proportion of the words in the response are at or above each CEFR proficiency level (A2, B1, B2, C1, C2), along with the proportion of words that fall outside the standard vocabulary altogether. Responses that draw on more advanced vocabulary receive higher values on the higher-level features. **Technical Implementation:** Each token in the response is lemmatized and looked up in an internal CEFR dictionary. Tokens found in the dictionary are considered "in-vocabulary"; those not found are "out-of-vocabulary" (*lex-comp-oov*). The *lex-comp-a2* through *lex-comp-c2* features are computed as the proportion of in-vocabulary lemmatized tokens at or above the specified CEFR level. **Relevant Citations:** Xia et al. (2019)
Differential Use of Vocabulary *du-token*, *du-lemma*	**High-level Description:** Similar to Differential Use of Grammar but for word usage. A response that uses words and word forms in ways characteristic of higher-proficiency writing receives a higher value, while one whose patterns resemble lower-proficiency writing receives a lower value. **Technical Implementation:** Identical to Differential Use of Grammar but defined over words and lemmas instead: *du-token* (3-gram model of raw tokens) and *du-lemma* (3-gram model of lemmatized tokens). **Relevant Citations:** Attali (2011); Yancey et al. (2021); Vajjala and Lučić (2018); Heafield et al. (2013)
Lexical Diversity *mtld* *lemma-mtld*	**High-level Description:** Measures the range of different words used in the response, distinguishing genuinely varied word choice from simple repetition of the same vocabulary. **Technical Implementation:** Measure of Textual Lexical Diversity (MTLD) computed over the raw tokens of the response (*mtld*) and over the lemmatized tokens of the response (*lemma-mtld*). MTLD reflects the mean length of sequential word strings that maintain a given type-token ratio threshold. **Relevant Citations:** McCarthy and Jarvis (2010)
Mean Word Length *mean-word-length*	**High-level Description:** Measures the average length of words in the response. Since longer words tend to be rarer and more advanced, this serves as a simple indicator of vocabulary sophistication. **Technical Implementation:** The mean character length of word tokens in the response. **Relevant Citations:** Attali and Burstein (2006)

Table 13 continued

Feature group *list of features*	**Details**
Vocabulary Control *learner-vocabulary-control*	**High-level Description:** Measures whether the test taker uses words in contexts that match how those words are used in writing by other English learners. Words used in typical contexts receive high values; words used in unusual or unexpected contexts receive low values. **Technical Implementation:** For each word in the response, a 6-gram model trained on a sample of high-scoring DET responses is used to compute the probability of the word appearing in its context. To avoid conflating contextual probability with raw word frequency, the feature incorporates the unigram probability of each word. The final feature is the log-ratio of the 6-gram probability to the unigram probability, averaged over all tokens in the response. **Relevant Citations:** Crossley et al. (2012)

Pronunciation

Table 14. Pronunciation features used to score speaking.

Feature group *list of features*	**Details**
Pronunciation Score *pronunciation-score*	**High-level Description:** A holistic score from a deep learning model trained to match human ratings of how clearly and intelligibly the test taker produces English sounds, word stress, and sentence intonation. The model is designed to align directly with three established dimensions of pronunciation proficiency: overall phonological control (general intelligibility), sound articulation (the production of individual consonants and vowels), and prosodic features (stress, rhythm, and intonation). **Technical Implementation:** A lightweight deep neural network pronunciation scoring model that takes as input intermediate audio representations from the ASR acoustic encoder. The model is trained on human pronunciation ratings aligned with three CEFR linguistic competence scales: overall phonological control, sound articulation, and prosodic features. Building the scorer on top of a shared acoustic encoder provides a strong acoustic representation of the response without requiring a separate large acoustic model, while the lightweight scoring head is trained specifically to predict construct-aligned human ratings rather than transcription accuracy. **Relevant Citations:** Cai et al. (2025)
Prosodic Features *pitch_normalized_range*, *energy_sd*	**High-level Description:** Measures the rhythm and melody of the response, i.e., how the voice rises, falls, and varies in emphasis across speech. These features capture the stress and intonation patterns that contribute to intelligible speech. These features complement the comprehensive pronunciation score by providing additional, directly interpretable acoustic measurements of prosody. **Technical Implementation:** Two signal-processing-based acoustic features adapted from ETS's SpeechRater v5.0 engine (L. Chen et al., 2018). *pitch_normalized_range* is the range of normalized pitch across the spoken response. *energy_sd* is the standard deviation of speech energy. **Relevant Citations:** L. Chen et al. (2018)
Speech Recognition Confidence *asr-score*	**High-level Description:** Uses the ASR model's own confidence in its transcription as an indirect indicator of intelligibility. Speech that the ASR system can transcribe with high confidence tends to be easier for human listeners to understand as well, while speech that produces low-confidence transcripts is often less intelligible. This feature is not a *direct* measure of pronunciation, since ASR confidence is also affected by factors such as lexical sophistication and audio quality, but it provides a useful complement to the more construct-aligned pronunciation score described above. **Technical Implementation:** A confidence score derived from the ASR model when transcribing the spoken response. Higher values reflect speech that the model can transcribe with greater certainty. **Relevant Citations:** Qian et al. (2019); Zhang et al. (2019)

C Writing Penalties

In this section, we describe the penalties applied to writing responses on top of the scores produced by the ML scoring system. Penalties exist to ensure that responses produced in **bad faith** (off-topic, written in a non-English language, unintelligible or incoherent, lacking substantive content, or composed primarily of generic filler language) receive appropriately low scores regardless of the surface-level features the response may exhibit. Penalties fall into two categories:

1. **Automatic penalties**: These penalties are applied directly by automated systems without human intervention.
2. **Proctor-applied penalties**: For such penalties, an automated system raises a signal that is then confirmed or rejected by the human proctor reviewing the test session before any score adjustment is made. This design combines the scalability and consistency of automated detection with the experience and expertise of human proctors.

Table 15 describes all penalties applied to DET writing responses. When more than one penalty is triggered for the same response, the response receives the most severe of the applicable penalties. Penalties are applied as a *cap* on the score rather than a fixed deduction, meaning that a response whose score is already at or below the penalty threshold is unaffected—penalties only lower scores that would otherwise have been higher. All penalties cap the penalized response's score at the equivalent of 80 points on the 10–160 reported scale, except Gibberish / Non-English, which caps it at 10 points.

Table 15. Penalties applied to DET writing responses.

Penalty	Details	Writing task(s)	Mode
Gibberish / Non-English	Written responses that are unintelligible (e.g., random strings of characters or meaningless sequences of words) or that are written in a language other than English are automatically assigned the lowest possible score. Detecting such responses reliably is non-trivial, since short or low-proficiency English responses can superficially resemble unintelligible text, and short responses in some languages can superficially resemble English. We use an ensemble of three *independent* detection systems and apply the penalty only when the ensemble agrees: a publicly available XLM-RoBERTa model fine-tuned for language identification (Conneau et al., 2020), the *lingua* library (Stahl, 2023), and a vocabulary filter built on a predefined English word list from the Natural Language Toolkit (NLTK; Bird et al., 2009). This design reliably catches responses that are *clearly* non-English or otherwise unintelligible while reducing the risk of unfairly penalizing genuinely English responses, e.g., low-proficiency responses that superficially resemble unintelligible text or responses containing occasional borrowed words or expressions for which no close English equivalent exists.	All writing task types	Automatic

Table 15 continued

Penalty	Details	Writing task(s)	Mode
Low content rating	Responses that contain very little substantive content relative to the prompt are detected using the LLM-based content rating feature (Appendix B, Table 9), which uses a fine-tuned GPT-4.1-mini model to assign each response an integer rating for how substantively its content addresses the prompt. The penalty is applied when this rating is either 0 or 1.	*Writing Sample*; *Interactive Writing (Initial)*; *Interactive Writing (Follow Up)*; *Write About the Photo*	Automatic
High nonsense fraction	A separate penalty targets responses composed largely of nonsensical text: strings of words that are individually plausible but do not combine into any coherent meaning. Detection relies on two LLM-derived signals acting in conjunction. The first identifies spans within the response—using a fine-tuned GPT-4.1-mini model—that are nonsensical and computes the fraction of the response (by character length) taken up by such spans. The second is the same content rating feature used in the Low content rating penalty above. The penalty is triggered only when the nonsense fraction ≥ 0.85 (i.e., the great majority of the response is nonsensical), *and* the content rating judges the response to be weak in substantive content. Requiring both these signals prevents false positives where the nonsense detector might flag long spans in a response that is otherwise reasonable.	*Writing Sample*; *Interactive Writing (Initial)*	Automatic

Table 15 continued

Penalty	Details	Writing task(s)	Mode
Off-topic	Responses that are irrelevant or only tangentially relevant to the prompt are detected using two independent signals, both computed for every response. The first is a topicness score produced by a ModernBERT-base classifier (Warner et al., 2025) fine-tuned on a large sample of historical responses labeled for off-topicness by GPT models, distilling their judgments into a lightweight model that can be applied at scale; the **topicness score** reflects how closely the response relates to the prompt. The second is a binary off-topic judgment from a separate fine-tuned GPT-4.1-mini model trained on expert human ratings of topicness. These two signals are combined in two different ways. The penalty is applied *automatically* when *both* signals agree that the response is off-topic and the topicness score falls below a strict threshold (0.25). Separately, the response is flagged for proctor review whenever *either* of the two inputs flags the response, with ModernBERT using a broader topicness score threshold in this case (0.4); if the proctor accepts the signal, the penalty is applied. Requiring *both* signals to agree before a penalty is applied catches the clearly off-topic cases while the more inclusive criterion for proctor review ensures that borderline cases still reach a human.	*Writing Sample*; *Interactive Writing (Initial)*	Automatic, Proctor-applied

Table 15 continued

Penalty	Details	Writing task(s)	Mode
Overuse of shell language	A second proctor-applied penalty handles responses with *shell language* overuse (Madnani et al., 2012; Mulcaire & Madnani, 2025). Such responses are highly templated or composed primarily of generic phrases of the following type without meaningfully responding to the prompt: • linking expressions • descriptive statements about the response itself • generic statements about the importance of the topic • restatements of the prompt • vague references to studies • stance-taking phrases Some amount of shell language is expected and even helpful in a well-written response. Therefore, we penalize only the responses that are *overwhelmingly* composed of such material at the expense of substantive content. The shell overuse signal is raised when two conditions are jointly satisfied: (i) a fine-tuned GPT-4.1-mini model trained on human ratings of shell overuse classifies the response as such and (ii) the LLM-based content rating feature (Appendix B, Table 9) is at or below a low threshold, providing independent corroboration that the response is weak in content. Responses meeting these two conditions are surfaced to proctors who then accept or reject the signal. As with the off-topic penalty, the penalty is applied if the signal is accepted.	*Writing Sample*; *Interactive Writing (Initial)*	Proctor-applied

D References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). GPT-4 technical report. https://doi.org/10.48550/arXiv.2303.08774

Attali, Y. (2011). A differential word use measure for content analysis in automated essay scoring. *ETS Research Report Series*, *2011*(2), i–19. https://doi.org/10.1002/j.2333-8504.2011.tb02272.x

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Beigman Klebanov, B., & Madnani, N. (2022). *Automated essay scoring*. Springer. https://doi.org/10.1007/978-3-031-02182-4

Belzak, W., Baig, B., Cardwell, R., Hastings, R., Horie, A., LaFlair, G., Liao, M., Niu, C., & Chih, Y. (2025). *Duolingo English Test: Security and score integrity* (Duolingo Research Report No. DRR-25-01). https://duolingo-papers.s3.us-east-1.amazonaws.com/reports/DET_Security_Report.pdf

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. O'Reilly Media.

Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. *Proceedings of the 55th annual meeting of the Association for Computational Linguistics*, 793–805. https://doi.org/10.18653/v1/P17-1074

Burstein, J. (2026a, March). *The Duolingo English Test responsible AI standards* (Duolingo Research Report No. DRR-26-02). Duolingo. https://go.duolingo.com/ResponsibleAI

Burstein, J. (2026b, August). *The Duolingo English Test responsible AI transparency report* (Duolingo Research Report No. DRR-26-08). Duolingo. https://go.duolingo.com/DET_RAI_Transparency_Report

Cai, D., Naismith, B., Kostromitina, M., Teng, Z., Yancey, K. P., & LaFlair, G. T. (2025). Developing an automatic pronunciation scorer: Aligning speech evaluation models and applied linguistics constructs. *Language Learning*, *75*, 170–203. https://doi.org/10.1111/lang.70000

Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., & Gyawali, B. (2018). Automated scoring of nonnative speech using the SpeechRaterSM v. 5.0 engine. *ETS Research Report Series*, *2018*(1), 1–31. https://doi.org/10.1002/ets2.12198

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

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: Learning, teaching, assessment – companion volume*. https://www.coe.int/lang-cefr

Crossley, S., Cai, Z., & McNamara, D. (2012). Syntagmatic, paradigmatic, and automatic *n*-gram approaches to assessing essay quality. *Proceedings of the 25th international FLAIRS conference*, 214–219.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., & Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. *Proceedings of workshop for NLP open source software (NLP-OSS)*, 1–6. https://doi.org/10.18653/v1/W18-2501

Goodwin, S., Attali, Y., LaFlair, G. T., Park, Y., Runge, A., von Davier, A. A., Yancey, K. P., Naismith, B., & Chuang, P.-L. (2025). *Duolingo English Test: Writing construct* (Duolingo Research Report No. DRR-22-03). Duolingo. https://go.duolingo.com/scored-writing

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 2: Short papers)*, 690–696.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python*. https://doi.org/10.5281/zenodo.1212303

Ikeda, N., Clark, T., Papageorgiou, S., Gu, L., Ohta, R., Blackhurst, A., & Bruce, E. (2025). *Aligning scores of language proficiency tests: A score concordance study between IELTS Academic and TOEFL iBT®* (TOEFL Research Report No. RR-105). ETS. https://doi.org/10.64634/tn27a897

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing* (2nd ed.). Prentice Hall.

Lee, K., He, L., & Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 2 (short papers)*, 687–692. https://doi.org/10.18653/v1/N18-2108

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742. https://doi.org/10.1162/tacl_a_00343

Lottridge, S., Godek, B., Jafari, A., & Patel, M. (2020). *Comparing the robustness of deep learning and classical automated scoring approaches to gaming strategies*. Cambium Assessment. https://files.portal.cambiumast.com/corporate-site/documents/CAI-Cambium-Comparing-Robustness-Automated-Scoring-Approaches.pdf

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765–4774, Vol. 30).

Madnani, N., Heilman, M., Tetreault, J., & Chodorow, M. (2012). Identifying high-level organizational elements in argumentative discourse. *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 20–28.

Marone, M., Weller, O., Fleshman, W., Yang, E., Lawrie, D., & Van Durme, B. (2025). mmBERT: A modern multilingual encoder with annealed language learning. https://doi.org/10.48550/arXiv.2509.06888

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In *Applied natural language processing: Identification, investigation and resolution* (pp. 188–205). IGI Global. https://doi.org/10.4018/978-1-60960-741-8.ch011

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. https://doi.org/10.48550/arXiv.1712.09405

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*, 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Mulcaire, P., & Madnani, N. (2025). Span labeling with large language models: Shell vs. meat. *Proceedings of the Twentieth Workshop on Innovative Use of NLP for Building Educational Applications*.

Munro, R. (2021). *Human-in-the-loop machine learning: Active learning and annotation for human-centered AI*. Manning Publications.

Naismith, B., Cai, D., Kostromitina, M., & LaFlair, G. T. (under review). *Validating automated measures of fluency for use in a high-stakes English proficiency test* [Manuscript submitted for publication].

Naismith, B., Cardwell, R. L., LaFlair, G. T., Nydick, S. W., & Kostromitina, M. (2026, July). *Duolingo English Test: Technical manual* (Duolingo Research Report). Duolingo. https://go.duolingo.com/dettechnicalmanual

Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA)*, 394–403. https://doi.org/10.18653/v1/2023.bea-1.32

Nydick, S. W., & Lockwood, J. R. (2025). *An overview of Duolingo English Test administration and scoring* (Duolingo Research Report No. DRR-24-03). Duolingo. https://go.duolingo.com/Admin+Scoring

OpenAI. (2025a). ChatGPT [Large language model]. https://chatgpt.com

OpenAI. (2025b, April). Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/

Park, Y., Attali, Y., Zhang, X., Church, J., Lo, K.-L., Runge, A., Goodwin, S., & LaFlair, G. T. (2026). Toward standardized interactivity: An AI-enabled adaptive speaking task for computer-delivered large-scale assessment [Advance online publication]. *Language Testing*. https://doi.org/10.1177/02655322261462124

Park, Y., Cardwell, R., Goodwin, S., Naismith, B., LaFlair, G. T., Yancey, K. P., & Zhang, X. (2025). *Assessing speaking on the Duolingo English Test* (Duolingo Research Report No. DRR-23-03). Duolingo. https://go.duolingo.com/speaking-whitepaper

Qian, Y., Lange, P., & Evanini, K. (2019). Automatic speech recognition for automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 61–74). Routledge.

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® scoring engine. *ETS Research Report Series*, *2009*(1), i–35. https://doi.org/10.1002/j.2333-8504.2009.tb02158.x

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th international conference on machine learning*, 28492–28518.

Rei, M., & Cummins, R. (2016). Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. https://doi.org/10.48550/arXiv.1606.03144

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 3982–3992. https://doi.org/10.18653/v1/D19-1410

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*, 523–530. https://doi.org/10.3115/1219840.1219905

Sekoyan, M., Koluguri, N. R., Tadevosyan, N., Zelasko, P., Bartley, T., Karpov, N., Balam, J., & Ginsburg, B. (2025). Canary-1B-v2 and Parakeet-TDT-0.6B-v3: Efficient and high-performance models for multilingual ASR and AST. https://doi.org/10.48550/arXiv.2509.14128

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and future directions*. Routledge.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Stahl, P. M. (2023). *Lingua*. https://github.com/pemistahl/lingua

Vajjala, S., & Lučić, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. *Proceedings of the 13th workshop on innovative use of NLP for building educational applications*, 297–304. https://doi.org/10.18653/v1/W18-0535

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Adams, G. T., Howard, J., & Poli, I. (2025). Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Proceedings of the 63rd annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, 2526–2547.

Weir, C. J. (2005). Language testing and validation. *Hampshire: Palgrave McMillan*, *10*, 9780230514577.

Williamson, J. (2026, January). *Principles of AI use in marking*. Office of Qualifications and Examinations Regulation (Ofqual). https://www.gov.uk/government/publications/principles-of-ai-use-in-marking/principles-of-ai-use-in-marking

Xia, M., Kochmar, E., & Briscoe, T. (2019). Text readability assessment for second language learners. https://doi.org/10.48550/arXiv.1906.07580

Yan, X., Lei, Y., & Pan, Y. (2025). Diving deep into the relationship between speech fluency and second language proficiency: A meta-analysis. *Language Learning*, *75*, 1051–1090. https://doi.org/10.1111/lang.12701

Yancey, K. P., Pintard, A., & François, T. (2021). Investigating readability of French as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e Linguaggio*, *20*(2), 229–258.

Zechner, K., & Evanini, K. (Eds.). (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge. https://doi.org/10.4324/9781315165103

Zhang, M., Yao, L., Haberman, S. J., & Dorans, N. J. (2019). Assessing scoring accuracy and assessment accuracy for spoken responses: Using human and machine scores. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 32–58). Routledge.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1–38. https://doi.org/10.1145/3639372