An Overview of Duolingo English Test Administration and Scoring

Duolingo Research Report DRR-24-03
October 2, 2024 (7 pages)
https://englishtest.duolingo.com/research

Steven W. Nydick and J.R. Lockwood

Abstract

In this report, we describe test administration on the Duolingo English Test (DET). The DET contains multiple test sections, each of which includes specific methods of choosing and delivering test items. There are two item administration methods, one based on choosing items given responses to earlier items, and the other randomly selecting items without reference to any response. Every computerized adaptive testing (CAT) item might depend on responses to all earlier CAT items. These administration methods map to two different scoring methods, one using item response theory (IRT) estimation and the other based on simple aggregates of item grades. This report describes the mechanics of the test administration/delivery and scoring methods as of July 1, 2024.

Last Update: October 2024

Keywords

Duolingo English Test, computerized adaptive testing, assessment, test administration, test scoring

Contents

Email: englishtest-research@duolingo.com

1 Introduction

The Duolingo English Test (DET) is a high-stakes, English language proficiency (ELP) test designed to support decisions, such as admission to a postsecondary institution, in English-medium settings. The DET was conceptualized as a digital-first, entirely online test, administered in approximately an hour with a variety of item types. To create a relatively short online test without compromising the reliability or security of the test, the DET contains a very large item bank with novel test administration methods. Given the complexity of administration and size of the item bank, very few items overlap on any random pair of tests. This document will provide a high-level overview of the test administration and scoring of the DET, including summarizing the sections of the test, describing how individual section scores influence the items selected in subsequent sections, and explaining how final scores are calculated. All information relates to DET test administration and scoring as of July 1, 2024. Many of the descriptions also apply to earlier versions of the DET to some degree but are not guaranteed to entirely reflect how those versions were administered.

2 Test Administration

The following section contains a brief description of test administration on the DET. More information about the item types is contained within the DET technical manual (Cardwell et al., 2024). Many papers and books describe the logic of adaptive testing, the process behind adaptive testing, and the justification for using adaptive methodology in operational testing programs (Thompson, 2009, 2011; van der Linden & Glas, 2010; Weiss, 1982). Given the plethora of resources available, we only provide a general overview of typical adaptive testing methodology.

The current DET contains adaptive and non-adaptive sections. An adaptive test works by choosing subsequent items based on information from prior items. Typically, if a test taker answers a question correctly, the next question will be a bit more difficult, and if a test taker answers a question incorrectly, the next question will be a bit easier. Each test taker will thus see a different set of questions, with test takers who answer earlier questions correctly receiving more difficult later questions than test takers who answer earlier questions incorrectly. Because test takers have systematically different tests, tests are typically scored by a complicated algorithm rather than simple number correct scores (which would be biased without considering the difficulty of the questions). Essentially, getting a question correct will increase the test taker's score, although the magnitude of this increase depends on the difficulty of the question. Moreover, getting a question incorrect will decrease the test taker's score with the magnitude of the decrease depending on the easiness of the question. The information attached to an item (such as how difficult an item is) is estimated from prior samples of data and is assumed to apply to all test takers. This assumption—that we can obtain global information about items from prior responses and use that information to predict future responses—is the critical assumption that underpins adaptive testing. Note that the mechanics behind this scoring process are beyond the scope of this document, although see Hambleton and Jones (1993) for a comparison of classical scoring methods with modern scoring methods and Embretson and Reise (2013) for more information as to the models used, how to estimate information about those models, and how to use that information to score typical assessments. We discuss and differentiate between these scoring methods in the following section.

Figure 1 provides a diagram as to the administration process of the DET. The diagram can be interpreted as follows. Each block is represented by a vertical rectangle. All items within a block are administered together prior to the next block. Task types within a block are not always administered together unless the block contains only a single item type. The color of the block indicates the type of section: orange for adaptive and green for non-adaptive. The total number of items contained within a block is provided below the rectangle. If multiple task types are contained within a single block, the number of items is broken out by task type. The arrow connecting two blocks indicates that information from the preceding block (the tail of the arrow) impacts items chosen in the subsequent block (the head of the arrow) as well as all future blocks that can be traced back to the first block by arrows. Individual skills (i.e., speaking, writing, reading, and listening) are typically highly correlated (e.g., Sawaki & Sinharay, 2013, for the TOEFL). Therefore, responses to questions testing one skill (e.g., reading) will provide information about how the test taker might perform on a different but related skill (e.g., listening). Finally, the DET algorithm chooses a reasonable set of questions based on the test taker's responses to earlier questions and then randomly selects one of the questions from that set. Adding a random component to item selection partially mitigates the tendency of optimal adaptive selection algorithms to administer the best items too often, which can compromise the security of the test.

At the beginning of the DET, test takers are presented with 15-18 Single Text Yes/No Vocabulary items, with each item chosen adaptively. That is, the first item in the test will typically be of moderate difficulty, test takers who get that item correct will typically be presented with a more difficult vocab item, and test takers who get that item incorrect will typically be presented with an easier vocab item.* Additionally, across all test sessions, approximately 50% of the presented vocab items will be real words, and approximately 50% of the presented vocab items will be plausible but fake words, although the exact proportion will vary across test sessions.

*Technically, the best initial guess of a person's ability is the average ability across all persons. This guess would lead to an initial item of moderate difficulty. However, because of the partial randomness of the algorithm, the actual chosen item might be more difficult or easier than this guess.

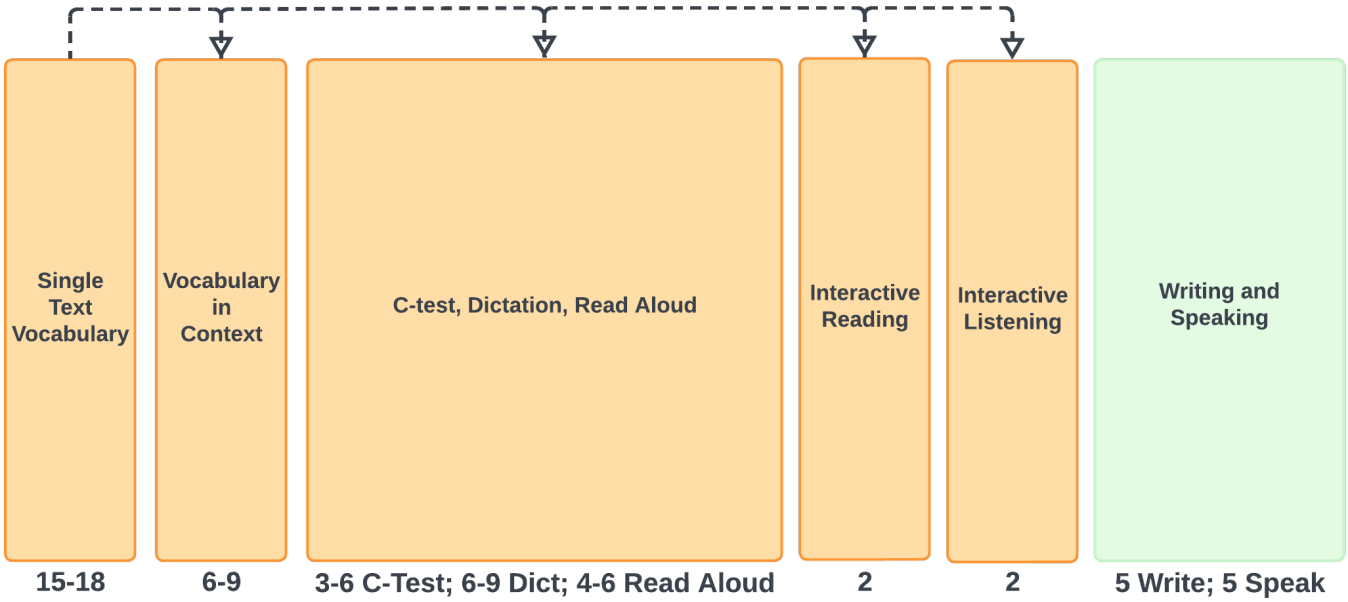

Figure 1. Simple diagram of current DET test administration sections. Items within a block are administered together within that block. Orange sections have adaptive administration, and green sections have non-adaptive administration. Text below a block indicates the approximate number of items administered in that block. Dotted arrows connecting two blocks indicates that information from the preceding block potentially impacts items chosen in the subsequent block.

After test takers complete the Yes/No Vocabulary section, they will be presented with 6-9 Vocabulary in Context (i.e., fill-in-the-blank) questions. The first Vocabulary in Context item might be chosen, in part, based on how well the test taker performed on the Single Text Yes/No Vocabulary section. Therefore, if test takers answered more Yes/No Vocabulary items correctly, they might start with more difficult Vocabulary in Context questions. Subsequent questions are chosen adaptively based on responses to scores on all earlier items on the CAT.

The third section of the test contains three task types: C-test, Dictation, and Read Aloud. Tasks in this section alternate such that the next question is chosen from the task type that has the most required remaining questions. For example, the first task type of this section is always Dictation because more Dictation questions are required than C-test or Read Aloud questions. Each item is chosen based on scores to all earlier items in the CAT.

The final two sections of the CAT portion of the test are the Interactive Reading and Interactive Listening sections, each of which contains two items. An Interactive Reading item includes a set of tasks based on a reading passage. An Interactive Listening item includes a set of tasks sequentially choosing the best statement or question to continue a conversation (usually referred to as a conversation "turn") as well as a final task to summarize the entire conversation (LaFlair et al., 2023). Each task (such as choosing the title or main idea for Interactive Reading or an individual turn for Interactive Listening) has its own difficulty, although the reading passages or conversations are chosen as a whole based on how difficult the item is in aggregate. Note that at present, the summarization task for Interactive Listening does not have an associated difficulty and thus does not influence item selection. The Interactive Reading items are first chosen based on the same process as the rest of the CAT sections, so that each item is picked based on responses to all prior items on the test. After completing the Interactive Reading section, both Interactive Listening items are chosen simultaneously based on responses to all preceding items.

The final section of the test includes three writing tasks where test takers are asked to describe an image; an Interactive Writing question where test takers respond to two writing prompts, the second prompt selected based on responses to the first prompt; several short speaking prompts (speaking about a photo, speaking about an auditory prompt, and speaking about a written prompt); a final Extended Writing prompt; and a final Extended Speaking prompt. The last two prompts are differentiated from the earlier writing and speaking prompts by requiring a longer response and sending that response to institutions along with grading that response as part of the final DET score. Each of these writing and speaking tasks is randomly selected without reference to earlier portions of the test. Only the Interactive Writing question includes a prompt that depends on an earlier response, but that earlier response is the first part of the same question, and the follow-up prompt is selected based on the content of the

first response rather than the grade. See Goodwin et al. (2022) and Park et al. (2023) for more information about how the DET assesses writing and speaking, respectively.*

3 Test Scoring

We can differentiate between two aspects of obtaining a final score on the assessment: grading and scoring. Grading is the process of turning an individual item response into one or more numbers capturing how well the individual responded to that item. Scoring takes those numbers and combines them, often using information from the item itself, such as its difficulty, to construct a score reported to test takers. This grading and scoring process also happens as part of test administration for the purpose of choosing subsequent items on the adaptive test, although the mechanisms behind how items are selected on the DET are slightly different from how the reported scores are generated. This section will focus on scoring for the purpose of reporting, although much of the information will also apply to the section on test administration. See Embretson and Reise (2013) for more details on the models used and van der Linden (2019) for the latest research on the models and their applications. Moreover, for the purposes of this section, we refer to a weighted combination of grades to determine the score as the classical method and model-based methods of scoring as the IRT method.

All items are automatically graded following the test-taker response. More information about the item response types are provided in the DET technical manual (Cardwell et al., 2024), and more information about the different grading engines are available in several white papers (Attali et al., 2022; Goodwin et al., 2022; LaFlair et al., 2023; Park et al., 2022, 2023, 2024). The following ranges are associated with grades on DET tasks.

1. Binary (i.e., 0 or 1, where 1 is correct and 0 is incorrect)
    1. Single Text Vocabulary
    2. Vocabulary-in-Context
    3. C-test (each damaged word)
    4. Interactive Reading (each Cloze Task blank, Text Completion, Select the Idea, Select the Title)
    5. Interactive Listening (each turn)
2. Continuous (between 0 and 1) with a point mass on 1 if fully correct
    1. Dictation
    2. Interactive Reading (Highlight)
3. Continuous between (approximately) 0 and 1
    1. Read Aloud
4. Continuous between (approximately) -3 and 3
    1. Writing (Writing Image, Interactive Writing, Extended Writing)
    2. Speaking (Speaking Image, Speaking Listen, Speaking Read, Extended Speaking)
    3. Interactive Listening (Summarization)

The mechanism behind grading depends on the task. Many of the tasks (e.g., Writing, Speaking, Read Aloud) use a complex algorithm to translate responses to grades using models that extract and process relevant information (Goodwin et al., 2022; Park et al., 2023). Other tasks (e.g., all binary response tasks) depend only on a simple model. The exact system for how responses are converted into grades is beyond the scope of this paper.

Ultimately, the grades that test takers receive on all items must be turned into final reported scores. The DET contains two different scoring processes depending on the response type. The processes mostly overlap with the steps of the adaptive algorithm, although this mapping need not always hold. Non-adaptive sections typically contribute to scores by a weighted combination of the grades. Adaptive sections typically result in scores by applying Item Response Theory (IRT) models (Embretson & Reise, 2013) that take into consideration the set of grades as well as information about the associated items, such as their difficulties. Note that IRT scoring does not create simple aggregates of item grades but pools information together across all relevant items and estimates the score that makes the most sense given those grades, information about the items, and the underlying models. Technically, we derive scores from IRT models by using the Expected A Posteriori (EAP) estimate given the item grades and the underlying IRT models, although the exact procedure is well documented in psychometric literature (e.g., Kim & Nicewander, 1993) and will not be discussed further here.

Figure 2 provides a diagram as to how grades contribute to final scores. As before, each block of items is represented by a rectangle. All items within a block are pooled together to obtain a preliminary score. The color of the block indicates the scoring process: red for the classical method, and blue for the IRT method. A line connecting two blocks indicates that the preliminary scores from those blocks are aggregated to form a composite score.

*All image-based tasks can elicit responses at all ability levels.

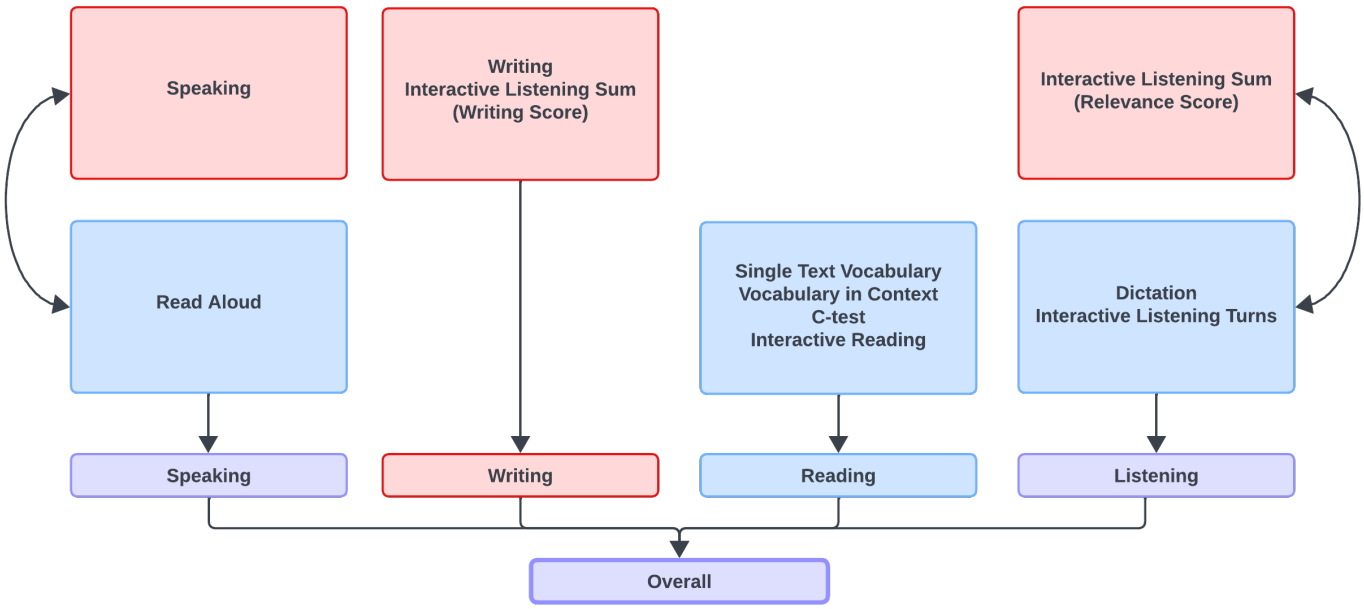

Figure 2. Simple diagram of current DET scoring process. Items within a block are scored together. Scores that consist of multiple blocks are scored by a weighted average of the scores on each of the blocks. Red blocks are scored by aggregating grades from the writing or speaking scoring engine. Blue blocks are scored by an Item Response Theory ability estimate given all grades within that block. Composite scores are color coded according to the component blocks: blue if composed of only blue blocks, red if composed of only red blocks, and purple if composed of blue and red blocks.

As of July 1, 2024, the DET contains four composite scores, which are then averaged to form an overall score on the test. These composite scores depend on six blocks, three with classical scoring and three with IRT scoring. Note that because the first four composite scores contain scores from a subset of tasks on the test, we typically refer to them as subscores (Cardwell et al., 2024).

The left-most subscore in Figure 2 represents speaking ability. This score is formed from an aggregation of two blocks: classically-scored speaking and IRT-based Read Aloud. A score on the classical speaking block is calculated by taking the speaking grades on each item (Speaking Image, Speaking Listen, Speaking Read, and Extended Speaking) and computing a weighted average. The IRT score is calculated by pooling all Read Aloud grades together and obtaining the EAP estimate given the appropriate IRT model. Overall speaking (i.e., the reported subscore) is a weighted average of the score on the classical block with the score on the IRT block.

A writing score is based only on a single block of writing items. The writing block is scored using classical methods, by taking a weighted average of relevant writing grades (Writing Image, Interactive Writing, Extended Writing and a summarization score for Interactive Listening). Because there are currently no IRT-scored items that contribute to the writing score, this aggregation then forms the final writing subscore.

The reading score depends only on an IRT block. To provide a reading score, all of the relevant item grades (Single Text Vocabulary, Vocabulary in Context, C-test, and all of the Interactive Reading items) are pooled together. Rather than calculating a score for each of the tasks and then aggregating those scores to form the reading score, all of the information is pooled together to obtain the EAP estimate given a combination of IRT models relevant to those tasks. This process essentially self-weights tasks by the information an individual item contains and the number of those items that a test taker answers. Note that for tasks with multiple grades per item (such as C-test and Interactive Reading), each part of a task, such as a single C-test blank or a single Interactive Reading response, is treated as a separate item for the purpose of scoring. Reading contains only IRT-scored items, so that the final reading score is taken directly from the EAP estimate of the IRT block.

A listening score is constructed from two blocks. The classical score is formed from the relevance features of the Interactive Listening summarization task. Unlike the Interactive Listening score that contributes to writing, this grade depends solely on the relevance of the response to the conversation and not on the quality or mechanical accuracy (such as using correct spelling rules) of a test taker's writing. An IRT score is calculated by pooling all grades from the dictation and Interactive Listening turns responses and estimating an EAP given the appropriate IRT models for each task. As with reading, responses to individual Interactive Listening turns are treated as separate items for the purposes of scoring. Listening is then scored by a weighted average of the score on the classical block and the score on the IRT block.

After constructing preliminary speaking, writing, reading, and listening scores, the algorithm converts each score to a DET scale score of 10-160 in 5-point increments for the purpose of reporting scores to institutions and test takers. These reported scores are then averaged to form a DET

overall score. As of July 1, 2024, pairs of these scores are also averaged to form interactive subscores, such as literacy (formed from reading and writing), conversation (formed from speaking and listening), comprehension (formed from reading and listening), and production (formed from speaking and writing). All scores are rounded to the nearest 5 for the purpose of reporting.

4 References

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & Von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation [Citation Key: attali2022]. *Frontiers in Artificial Intelligence*, *5*, 903077. https://doi.org/10.3389/frai.2022.903077

Cardwell, R., Naismith, B., LaFlair, G. T., & Nydick, S. W. (2024). *Duolingo english test: Technical manual* (tech. rep.) (Citation Key: cardwell-nd). https://duolingo-papers.s3.amazonaws.com/other/technical_manual.pdf

Embretson, S. E., & Reise, S. P. (2013, September 5). *Item response theory* (0th ed.) [DOI: 10.4324/9781410605269 Citation Key: embretson2013]. Psychology Press. https://doi.org/10.4324/9781410605269

Goodwin, S., Attali, Y., LaFlair, G. T., Park, Y., & Runge, A. (2022). *Duolingo english test-writing construct* (tech. rep.) (Citation Key: goodwin2022a). https://duolingo-papers.s3.amazonaws.com/other/writing-whitepaper.pdf

Hambleton, R. K., & Jones, R. W. (1993). An ncme instructional module on: Comparison of classical test theory and item response theory and their applications to test development [Citation Key: hambleton1993]. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests [Citation Key: kim1993]. *Psychometrika*, *58*(4), 587–599. https://doi.org/10.1007/BF02294829

LaFlair, G. T., Runge, A., Attali, Y., Park, Y., Church, J., & Goodwin, S. (2023). *Interactive listening–the duolingo english test* (tech. rep.) (Citation Key: laflair2023). https://duolingo-papers.s3.amazonaws.com/other/Interactive+Listening+%E2%80%93+The+Duolingo+English+Test.pdf

Park, Y., Cardwell, R., Goodwin, S., Naismith, B., LaFlair, G., Lo, K.-L., & Yancey, K. (2023, June 12). *Assessing speaking on the duolingo english test* (tech. rep.) (DOI: 10.46999/DJIY3654 Citation Key: park2023). https://doi.org/10.46999/DJIY3654

Park, Y., Cardwell, R., & Naismith, B. (2024). Assessing vocabulary on the duolingo english test [Citation Key: park2024].

Park, Y., LaFlair, G., Attali, Y., Runge, A., & Goodwin, S. (2022, June 16). *Interactive reading - the duolingo english test* (tech. rep.) (DOI: 10.46999/RCXB1889 Citation Key: park2022). https://doi.org/10.46999/RCXB1889

Sawaki, Y., & Sinharay, S. (2013, December). *Investigating the value of section scores for the toefl ibt test* (tech. rep.) (Citation Key: sawaki2013). https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02342.x

Thompson, N. A. (2009). Item selection in computerized classification testing [Citation Key: thompson2009]. *Educational and Psychological Measurement*, *69*(5), 778–793. https://doi.org/10.1177/0013164408324460

Thompson, N. A. (2011). Termination criteria for computerized classification testing [Publisher: University of Massachusetts Amherst Citation Key: thompson2011]. *Practical Assessment, Research, and Evaluation*, *16*, 1–7. https://doi.org/10.7275/WQ8M-ZK25

van der Linden, W. J. (Ed.). (2019). *Handbook of item response theory* [OCLC: 1124521537 Citation Key: vanderlinden2019]. Chapman & Hall/CRC.

van der Linden, W. J., & Glas, C. A. (Eds.). (2010). *Elements of adaptive testing* [DOI: 10.1007/978-0-387-85461-8 Citation Key: vanderlinden2010]. Springer New York. https://doi.org/10.1007/978-0-387-85461-8

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing [Citation Key: weiss1982]. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408