ets research institute

DECEMBER 2024

TOEFL® RESEARCH SERIES

Building a Validity Argument for the TOEFL Primary® Tests



Ching-Ni Hsieh

TOEFL–RR-103

ETS Research Report No. RR–24-16

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

Daniel F. McCaffrey
Lord Chair in Measurement and Statistics

ASSOCIATE EDITORS

Usama Ali
Senior Measurement Scientist

Beata Beigman Klebanov
Principal Research Scientist, Edusoft

Heather Buzick
Senior Research Scientist

Katherine Castellano
Managing Principal Research Scientist

Larry Davis
Director Research

Paul A. Jewsbury
Senior Measurement Scientist

Jamie Mikeska
Managing Senior Research Scientist

Teresa Ober
Research Scientist

Jonathan Schmidgall
Senior Research Scientist

Jesse Sparks
Managing Senior Research Scientist

Zuowei Wang
Measurement Scientist

Klaus Zechner
Senior Research Scientist

Jiyun Zu
Senior Measurement Scientist

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor/Communication Specialist

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Building a Validity Argument for the TOEFL Primary® Tests

Ching-Ni Hsieh

ETS Research Institute, ETS, Princeton, New Jersey, United States

Abstract

The TOEFL Primary® tests are designed to measure young language learners' language knowledge and skills to achieve core communication goals in the target language use domains of social-interpersonal and academic situations in primary and lower secondary education. This paper presents a fully articulated TOEFL Primary interpretive argument as a roadmap for the collection of evidence and evaluation of the TOEFL Primary validity argument. The validity argument comprises six validity inferences and a network of assumptions and claims about the measurement quality and the intended uses of the test scores. A considerable body of evidence was found to back up the claims that the TOEFL Primary tests are appropriate and useful for measuring students' English language proficiency, monitoring progress, and supporting teaching and learning. The paper provides score users and stakeholders with information about the test quality and can guide them to make decisions about the intended interpretations and uses of the test scores and their suitability for different learning contexts.

Keywords: English as an additional language, English language proficiency, TOEFL Primary®, young language learners, teaching, test scores, social-interpersonal, academics

Corresponding author: Ching-Ni Hsieh, Email: chsieh@ets.org

With the growing number of young students learning English as an additional language (EAL), the demand for high-quality, objective measures of English language proficiency (ELP) for young EAL learners has also increased (Leung, 2022; Wolf & Butler, 2017). To meet the needs of young EAL learners, ETS developed and introduced the TOEFL Primary® Reading and Listening tests and the TOEFL Primary Speaking test in 2013. In the ensuing years, score users had expressed interest in gaining insights into young EAL students' writing proficiency. In response to this need, ETS introduced the TOEFL Primary Writing test in 2023. Thus, the current TOEFL Primary tests include: (a) TOEFL Primary Reading and Listening Step 1 and Step 2 tests; (b) TOEFL Primary Speaking test; and (c) TOEFL Primary Writing test.

 © 2024 Educational Testing Service

The TOEFL Primary tests are designed to serve young EAL learners ages 8 and above. The tests measure test takers' knowledge, skills, and abilities (KSAs) to achieve core communication goals in English in the target language use (TLU) domains of social-interpersonal and academic situations in primary and lower secondary education. The tests can be used to (a) assess the English language abilities of young EAL learners to guide teaching and learning, (b) measure students' progress in attaining English proficiency, (c) inform placement decisions, and (d) evaluate performance against a widely accepted international standard (ETS, 2024; Wolf et al., 2024). The tests cover a wide range of ELP levels represented among young EAL students in different learning contexts and are independent of any specific language learning curriculum (ETS, 2019a). Detailed descriptions of the score reports and proficiency level descriptors can be found in the test-taker handbook (ETS, 2023a) and on the TOEFL Primary website: https://www.ets.org/toefl/primary/scoring-reporting.html.

TOEFL Primary Reading and Listening Tests

The TOEFL Primary Reading and Listening tests consist of a reading and a listening section and are delivered on paper or digitally. Each test section includes 36 scored questions and test takers have 30 minutes to complete the section. The test questions require test takers to read or listen to information, instructions, or stories that reflect young EAL learners' experiences at school, home, and other age-appropriate contexts. All questions are presented in a three-option, multiple-choice format.

The reading test measures young EAL learners' abilities to read a variety of written English texts in familiar contexts (e.g., school, home, playgrounds) and achieve the following communication goals:

- Identifying people, objects, and actions
- Understanding commonly occurring nonlinear written texts (e.g., signs, schedules)
- Understanding written directions and procedures
- Understanding short personal correspondence (e.g., letters, emails)
- Understanding simple, written stories
- Understanding written informational texts about age-appropriate academic topics

 © 2024 Educational Testing Service

The listening test measures young EAL learners' abilities to understand conversations related to daily life, spoken stories, and talks on familiar topics (e.g., animals). The test scores demonstrate test takers' abilities to fulfill the following communication goals:

- Understanding simple descriptors of familiar people and objects
- Understanding spoken directions and procedures (e.g., teacher directions in a classroom)
- Understanding conversations
- Understanding spoken stories
- Understanding short informational texts related to daily life (e.g., a voicemail message)
- Understanding expository monologues about age-appropriate academic topics

The TOEFL Primary Reading and Listening tests are offered at two difficulty levels—Step 1 and Step 2. Students can be matched with the test that best reflects their current level of proficiency. Step 1 is designed for young students at the beginning stages of learning English, and Step 2 is for students who have acquired some communicative skills in English.

The TOEFL Primary Reading and Listening tests are scored locally by ETS Preferred Network (EPN) offices. Scores are determined by the number of questions a test taker has answered correctly. The number of correct responses (i.e., the raw score) on each test section is converted to a scale score of 101–109 points for the Step 1 test and 104–115 points for the Step 2 test. The score report includes both a level and a scale score. Step 1 includes three levels which are shown as stars, and Step 2 includes four levels which are shown as badges. The score report also includes (a) proficiency descriptors of English language abilities typical of test takers scoring around a particular section scale score level, (b) the corresponding level on the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), and (c) a Lexile measure (MetaMetrics, 2024) based on the reading test section scale score. TOEFL Primary test takers can use the Lexile measures to find books targeted to their reading ability, grade level, or interest using the Lexile Find A Book tool (https://hub.lexile.com/find-a-book/; MetaMetrics, 2024).

TOEFL Primary Speaking Test

The TOEFL Primary Speaking test measures young EAL learners' abilities to communicate orally in routine social situations related to their daily lives. Test questions are administered via computer or tablet, and a headset is provided so students can listen to test

 © 2024 Educational Testing Service

prompts and record responses. The test scores demonstrate test takers' abilities to fulfill the following communication goals:

- Giving simple descriptions
- Expressing basic emotions, feelings, and opinions
- Making simple requests
- Asking questions
- Giving directions
- Explaining and sequencing simple events

The test tasks are presented in a scenario; animations and virtual characters (e.g., a zookeeper) appear at different times, functioning as interlocutors to create a communicative purpose, keep students engaged, and elicit spontaneous and natural responses. The test contains seven scored questions and five task types: Giving Directions, Narrating a Story, Asking Questions, Describing a Picture, and Expressing an Opinion. Some tasks incorporate scaffolding (e.g., key words) to support young students and enhance their test-taking experience. The test takes about 20 minutes to complete. The recorded spoken responses are scored by trained human raters using two separate scoring rubrics with either a 3-point or a 5-point rubric, depending on the task type (ETS, 2022a, 2023a). The total test score is reported on a scale of 1 to 27. The score report provides information about (a) an overall level (five total levels shown as ribbons), (b) proficiency level descriptors, (c) the total score (i.e., sum of the task raw scores), and (d) the corresponding CEFR level (ETS, 2023a).

TOEFL Primary Writing Test

The TOEFL Primary Writing test measures young EAL students' computer-based English writing abilities to communicate about familiar topics related to their daily lives. The test assesses a wide range of writing abilities, from writing words to writing a cohesive paragraph. The test questions are administered via computer or tablet, and a headset is provided so test takers can listen to the test prompts. The test scores demonstrate test takers' abilities to fulfill the following communication goals:

- Describing familiar objects, people, animals, places, activities, and situations
- Reading and making appropriate edits in order to make the text meaningful and accurate
- Sequencing simple events

 © 2024 Educational Testing Service

- Writing a short story

The writing test includes four task types: Write a Word (5 items), Build a Sentence (5 items), Edit a Text (4 items), and Write a Story (1 item). The test takes about 30 minutes to complete, including time for directions and practice questions. These tasks are situated in a scenario related to everyday or school-based situations and are sequenced from simple word- or phrasal-level production to more extended communication with increasing complexity (Wolf et al., 2024). The test score is reported on a scale of 0 to 17 and is the sum of the number of correct responses to the Write a Word, Build a Sentence, and Edit a Text tasks plus the score for the Write a Story task (0–3), which is produced using ETS's automated writing evaluation (AWE) engine trained on human ratings of students' responses (Li & Futagi, 2023). The score report provides information about (a) an overall level (four total levels shown as ribbons), (b) performance level descriptors, (c) the total score (i.e., sum of the task raw scores), and (d) the corresponding CEFR level.

Argument-Based Approach to Validation

The TOEFL Primary validity argument is built following an argument-based approach to validation (Kane, 2006, 2013, 2021) and draws on the TOEFL iBT and TOEFL Junior validity arguments (Chapelle et al., 2008; Hsieh, 2024). Kane stated that an argument-based approach to validation involves two main requirements: (a) a statement of the claims that are being made and (b) an evaluation of the credibility of these claims. Kane (1992) first introduced the argument-based validity framework using Toulmin's (2003) framework of argumentation, where claims about a test are supported by warrants (general statements used to justify a claim) or contradicted by rebuttals (alternative explanations or counterarguments to a claim). Warrants and/or rebuttals can be supported or refuted based on backing or evidence gathered from relevant documentation and theoretical or empirical research findings.

Kane (2006, 2013, 2021) expanded the original framework and identified two types of arguments: an *interpretive argument* and a *validity argument.* Kane explained that validation is an ongoing process that starts with outlining the proposed interpretations and uses of test scores in an interpretive argument and explicitly laying out the claims being made. An interpretive argument represents a network of inferences and assumptions that specify the proposed score interpretations and uses and conceptually links test performances to conclusions and decisions based on test scores. Once the interpretive argument is in place, test developers or researchers

 © 2024 Educational Testing Service

can then evaluate the coherence and plausibility of the proposed interpretations and uses in the interpretive argument. Simply put, the interpretive argument outlines the steps validation needs to go through and in what way, and the validity argument refers to how well evidence supports or challenges the interpretive argument.

The argument-based validation framework has the advantage of helping researchers identify the types of validation research required to support score interpretation and use, thereby providing links between the validity argument and validation research (Knoch & Chapelle, 2018). Such framework has been employed in a number of high-stakes language assessments such as the TOEFL iBT® (Chapelle et al., 2008) and TOEIC® (Schmidgall, 2017) tests and low-stakes tests such as an L2 pragmatic test (Youn, 2015) and an academic vocabulary test (Nguyen et al., 2024). Despite the clear trend for the growing use of the argument-based validation framework in language proficiency tests, this approach has not yet been a common practice in standardized language assessments designed for young students (e.g., Getman, 2020; Hsieh, 2024). Building on existing argument-based validation research, this paper presents the validity argument built for the TOEFL Primary tests.

TOEFL Primary Interpretive Argument

The development of a test and the creation of an interpretive argument are typically interwoven and ideally should be carried out in tandem (Kane, 2021). Within the context of the TOEFL Primary tests, during test development, test developers and researchers at ETS proposed a research agenda to support test design decisions and identify areas of future research with the goal of building a validity argument (Cho et al., 2016). To gather initial validity evidence supporting score interpretation and use, the test developers carried out multiple rounds of prototyping, pilot testing, and field studies to define the test constructs, evaluate the proposed assessment tasks and scoring procedures, finalize task design and test specifications, and create score reports. These development and validation efforts were firmly grounded within the argument-based approach to validation (Cho et al., 2016) and documented in several internal research reports and external publications. The test developers drew directly on the TOEFL iBT validation framework proposed by Chapelle et al. (2008) who formulated a series of warrants and assumptions in relation to six inferences (domain description, evaluation, generalization, explanation, extrapolation, and utilization) for the TOEFL iBT validity argument.

 © 2024 Educational Testing Service

Informed by previous work and drawing on the argument structure described in Chapelle et al. (2008), the current paper presents a fully articulated TOEFL Primary interpretive argument that serves as the basis for building a validity argument. A network of inferences, warrants, and assumptions that pertain to the TOEFL Primary test score interpretation and use is presented in Table 1.

Table 1. TOEFL Primary Interpretive Argument

Inference	Warrants	Assumptions
Domain description	TOEFL Primary test tasks represent relevant language knowledge and skills in situations representative of those in the TLU domain.	1. Assessment items and tasks produce evidence of language knowledge and skills representative of those used in the TLU domain. 2. Teachers perceive the test tasks as relevant and representative of those taught in the TLU domain.
Evaluation	Scores on the TOEFL Primary tests reflect the targeted KSAs.	1. Scoring procedures and guidelines are well specified and followed. 2. Scoring rubrics are appropriate for judging evidence of the targeted language knowledge and skills. 3. Statistical characteristics of test items and tasks are appropriate for providing evidence of the targeted language knowledge and skills. 4. Task administration conditions are appropriate for providing evidence of the targeted language knowledge and skills.
Generalization	Similar scores on the TOEFL Primary tests are expected to be obtained across comparable tasks, test forms, administrations, and testing conditions.	1. Sufficient numbers of items and tasks are included to provide reliable estimates of test-taker performance. 2. TOEFL Primary scores are consistent across test administrations. 3. Standard procedures are used to ensure score consistency across test forms. 4. Clearly defined item specifications are used to create parallel items and tasks.
Explanation	TOEFL Primary scores can be attributed to the constructs of social and academic ELP in primary and lower secondary-level education.	1. The strategies and processes required to successfully complete the TOEFL Primary tasks vary in keeping with theoretical expectations. 2. Task difficulty is systematically influenced by task characteristics.

 © 2024 Educational Testing Service

		3. Test performance varies according to experience in learning English.
Extrapolation	TOEFL Primary test scores reflect students' performances in the TLU domain in primary and lower secondary school contexts.	1. Performance on the TOEFL Primary tests is related to other criterion measures of ELP in the TLU domain.
Utilization	TOEFL Primary tests are useful for measuring students' language proficiency, monitoring progress, and supporting teaching and learning.	1. The meaning of test scores is clearly interpretable by test takers, teachers, parents, and other score users. 2. TOEFL Primary test scores are appropriate for the intended purposes.

Note. ELP = English language proficiency; KSAs = knowledge, skills, and abilities; TLU = target language use.

TOEFL Primary Validity Argument

The TOEFL Primary validity argument provides an overall evaluation of the interpretive argument to determine the degree to which the accumulative evidence supports the proposed test interpretations and uses. Kane (2021) stated that the purpose of a validity argument is to make explicit the meaning of a test score and the basis for claiming the various aspects of score meanings. The meanings of the TOEFL Primary test scores include the constructs of English language ability that they are supposed to represent, the degree of their consistency, their relationship with students' language use on other language tasks, and their appropriateness for the intended uses.

In this paper, validity evidence was collected from four main sources: (a) ETS's internal documents pertaining to standard operational procedures, (b) internal research memos and presentations that described or informed test development efforts, (c) published research papers and conference presentations, and (d) publicly available resources and documents. The collected evidence has been aligned to the assumptions and warrants underlying each inference of the interpretive argument and is discussed in the sections below. It should be noted that an AI tool was utilized to edit portions of the texts below to enhance readability.

Domain Description

The domain description inference is based on the warrant that the TOEFL Primary test tasks reflect the relevant KSAs in situations representative of those found in the target domain of language use in primary and lower secondary school contexts. This warrant is based on the

 © 2024 Educational Testing Service

assumptions that (a) the assessment tasks can produce evidence of the KSAs representative of those used in the TLU domain and (b) English language teachers perceive the test tasks as relevant and representative of those used in the TLU domain.

Assumption 1

Assessment items and tasks produce evidence of language knowledge and skills representative of those used in the TLU domain.

Backing 1. Domain analysis activities helped identify key assessment tasks that required the language knowledge and skills used in the TLU domain.

The TLU domain of TOEFL Primary tests includes social-interpersonal and school contexts where young EAL learners use the English language to communicate with people around them (e.g., family, friends, classmates, teachers) on daily life topics (e.g., people, animals, school) and obtain basic information from familiar sources (e.g., stories, directions; Cho et al., 2016). Backing for the assumption that test tasks reflect this TLU domain first came from domain analysis studies that focused on understanding language learning tasks in EAL classrooms. In an early study, Turkan and Alder (2011) reported the outcomes of a systematic review of national primary school English curricula and textbooks used in Brazil, Chile, China, Egypt, Japan, Korea, the Philippines, Qatar, and Singapore. The researchers identified common EAL classroom language tasks and learning objectives to define key communication goals and inform test design. During test development, test developers at ETS consulted the European Language Portfolios (ELPs) from various European countries to capture language proficiency standards for primary school students (Cho et al., 2016). Similarly, for the newly launched TOEFL Primary Writing test, the development team reviewed age-appropriate performance descriptors from the CEFR for young learners (Council of Europe, 2018a, 2018b) and English proficiency descriptors for school-aged L2 learners from Australia, Canada, and the United States (Wolf et al., 2024) to identify key writing communication goals for young learners.

Results of the multiple domain analysis efforts helped define the test constructs and determine critical communication goals unique to young EAL learners' communicative needs and the language knowledge and skills required to fulfill these communication goals. The communication goals subsequently informed the identification of specific text types that young learners encounter in typical English reading and listening materials as well as various types of speaking and writing activities in which young EAL learners engage in the classrooms.

 © 2024 Educational Testing Service

Backing 2. Empirical investigations confirmed the relevance and importance of the assessment tasks in the TLU domain for young EAL learners.

Results of two alignment studies provide support for the relevance and importance of the TOEFL Primary test tasks. Hsieh et al. (2013) conducted an alignment study to evaluate the match between the content of an EAL curriculum, Sesame Street English, and the TOEFL Primary tests. Sample instructional materials and lesson content from the Sesame Street English Primary 3 and Primary 4 courses were evaluated by the researchers for alignment with the TOEFL Primary TLU domain, task types, and communication goals. The analyses yielded varying degrees of alignment between the curricular activities and the TOEFL Primary test content. The courses' contents focused more heavily on the academic domain (76% of the tasks evaluated), while TOEFL Primary targeted both the social and academic domains. Findings revealed that the TOEFL Primary tests incorporated a wider range of task types and communication goals compared to those taught in the Sesame English Street courses, possibly due to the wider range of ELP skills targeted by the tests. In a second study, Papageorgiou et al. (2020) examined the degree of match between the TOEFL Primary tests and an online-delivered EAL curriculum, VIPKid Major Course. Three curriculum developers from the course provider were trained to systematically evaluate the content overlap between the TOEFL Primary tests and the curricular activities. Findings indicated overall good alignment of the language knowledge and skills, communication goals, and task types targeted by the TOEFL Primary tests and the Major Course content and class activities.

Further evidence was found in Chen and Sheehan (2015) who analyzed and compared the complexity of reading stimulus materials from the TOEFL Primary, TOEFL Junior®, and TOEFL iBT tests. Using an automated text analysis tool, TextEvaluator® (Sheehan et al., 2010), the researchers calculated text complexity scores for 17 TOEFL Primary reading passages that included academic reading, short reading, and narrative reading text types. The complexity scores compare the linguistic complexity of the reading passages to those of exemplar texts representative of reading materials used at different grade levels in the Common Core State Standards for English language arts (Common Core State Standards Initiative, 2010). The TOEFL Primary reading text complexity scores estimated all fell within the range of 0.36 and 8.40, a range that corresponds to reading standards specified for students in Grades 2 to 5 in U.S. primary schools.

 © 2024 Educational Testing Service

Assumption 2

Teachers perceive the test tasks as relevant and representative of those taught in the TLU domain.

Backing. Teachers thought that the TOEFL Primary test tasks were relevant to and representative of the language knowledge and skills taught in their English language classes or EAL curricula.

Hsieh (2013) explored EAL teachers' perceptions of the relevance and importance of the proposed TOEFL Primary communication goals and the appropriateness of the reading, listening, and speaking pilot tasks. The teachers were from Costa Rica (n = 4), Egypt (n = 5), Japan (n = 4), Peru (n = 8), and Vietnam (n = 8), teaching at local public, private, or international primary schools or after-school English language programs. They considered that the communication goals closely aligned with the communicative needs of their young EAL students. They also made recommendations regarding areas of misalignment such as the length of informational texts and the inclusion of an inappropriate task type. Based on these results, task refinement was undertaken to enhance the representativeness and appropriateness of the TOEFL Primary test tasks for young EAL learners.

Following test launch, Hsieh (2016) further confirmed teachers' perceptions of the content representativeness of the TOEFL Primary reading and listening test tasks. In this study, input was sought from 17 primary school teachers representing 15 countries. The teachers were provided with a test booklet that contained 57 reading items and 57 listening items. These items covered the full range of difficulty levels and task types assessed in the TOEFL Primary Reading and Listening Step 1 and Step 2 tests. Using a survey questionnaire, the teachers judged (a) the extent to which the content of the reading and listening test items reflected the target constructs and (b) the importance of the language knowledge and skills assessed by these items. Overall, the items were widely perceived as accurately reflecting the target constructs and effectively covering the essential domains of language knowledge and skills important for young EAL learners' communication needs.

Teacher perception of the TOEFL Primary test tasks was further examined as part of a washback study (Gu et al., 2016). In that study, nine classroom teachers from a Peruvian private language institute were recruited to assess the importance of the TOEFL Primary communication goals relative to their classroom activities and learning objectives. Most of the listening and

 © 2024 Educational Testing Service

speaking communication goals were perceived as critical and relevant for students' language development and language instruction. The teachers also reported using similar types of speaking and listening activities, such as "understanding instructions, directions, or procedures" and "understanding spoken descriptions" in their classrooms. However, the reading communication goal, "understanding expository or informational texts," was viewed as less important for younger learners between 7 and 8 years of age, which was the main age range of some teachers' students, though these more advanced texts were considered appropriate for older children.

Taken together, outcomes of these studies provide empirical evidence supporting the claim that the TOEFL Primary test tasks reflect relevant and important KSAs representative of those of the TLU domain of primary and lower secondary education.

Evaluation

The evaluation inference is based on the warrant that TOEFL Primary test scores reflect the targeted language abilities and skills. This warrant is dependent on the assumptions about scoring procedures, item analyses, and conditions of task administration. Backing in support of these assumptions includes evaluation of scoring criteria and procedures, implementation of rater training and calibration, statistical analyses of test items, and investigation of test administration conditions on test-taker performance.

Assumption 1

Scoring procedures and guidelines are well-specified and followed.

Backing 1. Standard procedures and guidelines are followed to score the TOEFL Primary Reading and Listening tests.

The procedures and guidelines for scoring the reading and listening test items and producing scaled section scores are detailed in an internal document that describes the procedures applied to the tests for item analysis, differential item functioning analysis, equating, and scaling (i.e., converting raw score points to scale scores; ETS, 2016a). These procedures are conducted and strictly followed by ETS's assessment developers, data analysts, and psychometricians who perform psychometric work for the TOEFL Primary Reading and Listening test administrations.

Backing 2. Human raters follow standard procedures and guidelines to score the TOEFL Primary Speaking test.

 © 2024 Educational Testing Service

The procedures and guidelines for scoring the TOEFL Primary Speaking responses are specified in a rater training manual (ETS, 2023b). The manual provides an overview of the test and task types and explains the scoring processes, scoring rubrics, and benchmark samples as well as rater calibration. Raters need to complete a training program and pass a certification test using the ETS Online Network for Evaluation system before they begin operational scoring (ETS, 2018a). The TOEFL Primary testing program regularly monitors the accuracy and reliability of scoring to ensure that raters apply the rubrics accurately and consistently (ETS, 2019b).

Assumption 2

Scoring rubrics are appropriate for judging evidence of the targeted language knowledge and skills.

Backing 1. Empirical research showed that raters were able to score reliably and identify differences in spoken performances across proficiency levels.

Getman (2020) detailed the process of developing and trialing the scoring rubrics for the TOEFL Primary Speaking test. In the study, a team of nine experienced, trained raters double-scored more than 400 students' spoken responses to the pilot speaking tasks. Evaluation of how well the test tasks, raters, and rating scales performed and discriminated among young EAL learners of varying speaking abilities was conducted through an array of different statistical analyses. The exact agreement rate between the raters was 74.0% for tasks scored using a 3-point scale and 59.4% for the tasks scored using a 5-point scale. The test takers significantly differed in ability with over eight statistically distinct groups. The reliability of separating the test takers based on their performance was high, with a coefficient of 0.98. The results suggested that the raters were able to reliably differentiate among young EAL students across a range of speaking ability levels.

Backing 2. Construct-relevant linguistic features are used to produce automatic scores for TOEFL Primary Writing test responses that are consistent with human scores.

Test-taker responses to the Write a Story task on the TOEFL Primary Writing test are scored using ETS's AWE engine. The AWE engine was trained on human ratings of 2,261 student responses collected from a field testing. The AWE engine covers the same scoring criteria used in the scoring rubrics, including aspects of content, cohesion and coherence, and language use. The agreement between AWE scores and human scores was evaluated prior to the

 © 2024 Educational Testing Service

implementation of automated scoring. Pearson correlation between machine and human score was 0.91, higher than the correlation between two human raters, which was 0.85 (Wolf et al., 2024; Wolf & Suhan, 2023). Human–machine score agreement was also evaluated across test takers who varied in terms of country of origin, first language, age, grade level, gender, and CEFR levels. Mean score differences across groups, based on test-taker backgrounds, revealed no meaningful differences between AWE and human scores (Li & Futagi, 2023). These results provide evidence supporting construct coverage and the consistency of machine scoring for the TOEFL Primary Writing test.

Assumption 3

Statistical characteristics of test items and tasks are appropriate for providing evidence of the targeted language knowledge and skills.

Backing. Empirical performance on test tasks was examined to verify that the test tasks were of appropriate difficulty and scores differentiated across levels of ability.

During test development, the TOEFL Primary test developers carried out a series of pilot studies to examine the influence of task features on the psychometric properties of the test tasks to inform test form compilation. For the TOEFL Primary Reading and Listening tests, pilot data were collected from over 1,300 students across eight countries. The analysis of these data yielded item difficulty and discrimination parameters as well as test form reliabilities (Cho et al., 2016, 2017). The results led to the decision to offer the TOEFL Primary Reading and Listening tests at two difficulty levels to ensure that the full range of students' reading and listening abilities was covered while maintaining test reliability. The pilot-test findings also helped determine the types of items to be included in each test level and test form to ensure adequate construct coverage.

The psychometric quality of the TOEFL Primary Reading and Listening test items was further supported by a field-test study conducted by Zu et al. (2017). In this study, six field-test forms, comprising 123 reading items and 120 listening items, were administered to over 3,700 students from 36 countries. The tests were administered in the students' classrooms to simulate operational testing conditions. The reading and listening test items exhibited a wide range of item difficulty and an appropriate level of discrimination. The average item discrimination estimate was 0.97 for the reading items and 0.92 for the listening items. These estimates suggest that most items effectively discriminated among the test takers' abilities.

 © 2024 Educational Testing Service

The psychometric quality of the TOEFL Primary Speaking test tasks was investigated in Getman (2020, 2024). Getman found that student ability measures based on pilot test scores were spread across a wide range, suggesting good discrimination across differing levels of ability of pilot test takers. The pilot test tasks appeared to cover a wide range of difficulty, demonstrating that they effectively assessed different proficiency levels. Furthermore, a differential facet functioning analysis indicated that the speaking tasks consistently functioned similarly for test takers across different age groups, and that no evidence of rater bias was observed in the evaluation of the speaking responses.

Additionally, ETS's data analysts and psychometricians continue to conduct routine analysis of item characteristics, following *ETS Standards for Quality and Fairness* (ETS, 2014). The TOEFL Primary testing program makes available brief reports on the distributions of test-taker demographic characteristics, average performance of groups of test takers, and percentile ranks of scale scores (ETS, 2022b).

Assumption 4

Task administration conditions are appropriate for providing evidence of the targeted language knowledge and skills.

Backing. Empirical studies showed that task administration conditions produced trustworthy evidence of ability.

Backing for the assumption that the TOEFL Primary task administration conditions are appropriate first came from prototyping studies described in the test design framework paper (Cho et al., 2016). For example, for the Speaking test, both computer-based and face-to-face individual and group administrations, were evaluated during test development. Trials of different formats focused on whether young EAL learners were able to perform the tasks in a manner that was appropriate and engaging given their age and proficiency levels. Insights gleaned from the trials revealed several issues related to the face-to-face administration format, such as children's lack of comfort when talking to adult examiners they had never met before and unequal assessment opportunities among young students in group or pair administrations. Informed by the results of the prototyping studies, the computer-delivered test format was adopted to ensure that students can demonstrate their true abilities under conditions that allow them to exhibit what they know.

 © 2024 Educational Testing Service

The appropriateness of the computer-mediated TOEFL Primary Speaking test was also examined in a nontesting research setting. Lee and Winke (2018) explored the response processes of 28 young children when they responded to TOEFL Primary speaking tasks on the computer. The researchers captured the children's eye movements using an eye-tracker during task performance. Results showed that low-proficiency non-native speaking (NNS) children viewed the timer more often and for longer periods compared to high-proficiency NNS children and native-speaking children. The researchers argued that the countdown timer could be a distracting test design feature because it introduced construct-irrelevant variance in processing and performance of some young children. This finding poses a rebuttal to the claim that the speaking task administration condition was appropriate for young learners.

The digitally delivered TOEFL Primary Reading test was examined in Ballard (2016). Participants were 31 children, between 8 and 10 years of age, with various first-language backgrounds. They responded to seven reading practice items on the computer and drew a picture to express how they felt during the test or what they thought about the test. They also participated in follow-up interviews and explained their drawings to the researcher. Findings showed that the computer-mediated testing format was not a hindrance to students' test performance and that children' attitudes toward the computer-delivered reading test were largely positive or neural, regardless of their familiarity with computer-mediated language assessments.

Test administration conditions of the TOEFL Primary Listening test were evaluated in two empirical studies. The first study, carried out by the test developers, focused on the evaluation of the impact of reading aloud and printing response options on paper-based listening test performance (Getman et al., 2016). The project team created different experimental test forms consisting of (a) a reading section, (b) a listening section with comprehension questions and response options printed on paper and also read aloud, and (c) a second listening section with comprehension questions and response options read aloud only. A sample of 747 EAL students responded to one of the test forms. Results indicated that presenting response options both aurally and in print did not statistically significantly impact the difficulty of the items for young learners. Student survey responses showed that close to 80% of the students preferred having the options available both in print and read aloud. The second study, conducted by an external team of researchers, examined the same issue using sample test items delivered on the computer (Hui et al., 2022). The researchers created two versions of sample TOEFL Primary

 © 2024 Educational Testing Service

listening conversation items: one fully read aloud and the other partially read aloud. Forty school-aged children in Hong Kong responded to the listening items under the two conditions. Findings showed that reading aloud the listening test items had the benefit of drawing students' attention to the content and guiding them through the test-taking process. Findings of the two studies collectively provide support for the appropriateness of presenting the TOEFL Primary listening test items both aurally and in print.

For the TOEFL Primary Writing test, the primary concern in test administration condition pertained to the use of a computer-based format (Wolf et al., 2024). Digital delivery of the writing tasks was initially chosen to allow the test tasks to be presented in scenario-based contexts, thereby enhancing engagement and promoting meaningful communication. The digital format could also facilitate the embedding of scaffolding and provide support for young learners during the writing tasks. Through a series of prototyping studies, critical test administration parameters such as the response format (e.g., constructed response, drag and drop), typing on a keyboard, and time allocated for each task were evaluated. These evaluation efforts contributed to the final task specifications and helped ensure effective task administration conditions for the TOEFL Primary Writing test.

Generalization

The generalization inference is based on the warrant that similar test scores are expected to be obtained across comparable test tasks, forms, administrations, and testing conditions. Four assumptions underlie this warrant, concerning issues related to composition of test forms, scaling and equating procedures, as well as the use of item specifications. Backing of these assumptions was obtained through the task design process and studies of reliability, scaling, and equating.

Assumption 1

Sufficient numbers of items and tasks are included to provide reliable estimates of test-taker performance.

Backing 1. Pilot- and field-testing studies helped determine the optimal number of TOEFL Primary test tasks needed in each test form to ensure score reliability.

The TOEFL Primary test developers carried out a series of pilot- and field-testing studies throughout test development to determine suitable task types and appropriate test form composition to measure the intended test constructs and maximize score consistency. For example, during pilot testing of the TOEFL Primary Reading and Listening tests, one key

 © 2024 Educational Testing Service

research question was to determine the number of test questions required in each test form to ensure score reliability. The analysis of the pilot-test data, involving responses from more than 1,300 students, helped finalize the length of the reading test and the distribution of questions and item types across the two levels (Step 1 and Step 2). The pilot-testing studies also showed that young EAL students' performances were impacted by the length of the tests since students' performances declined in longer test forms (Cho et al., 2016). Given the unique characteristics of young learners and the constraints of testing time, the optimal numbers of test questions were determined within a testing time of 30 minutes for each test section while maintaining test reliability for intended interpretations (Cho et al., 2016, 2017; Zu et al., 2017).

For the TOEFL Primary Speaking test, issues related to test form configuration and score reliability were addressed in Getman (2020, 2024). The researcher explored the impact of varying the number of tasks, ranging from five to ten, and the inclusion of one or two tasks for each task type on the reliability of the total test score. Based on a series of statistical analyses and consideration of the appropriate test length for young learners, a final test form configuration was established to maximize score reliability, involving seven tasks across five task types for each test form. Further analyses confirmed that the contributions of different task types to score reliability remained consistent across learners of different age groups. These research findings reinforce the assumption that TOEFL Primary test scores serve as reliable estimates of expected performances across various test forms and administrations for young EAL learners.

Backing 2. TOEFL Primary tests scores are internally consistent.

The internal consistency reliability estimates based on test-taker performances and section scale scores are .83 and .87 for TOEFL Primary Reading Step 1 and Step 2 tests, respectively; .85 and .89 for Listening Step 1 and Step 2 tests, respectively; and .90 for the Speaking test (ETS, 2019c, p. 29). These estimates indicate high levels of reliability for the low-stakes TOEFL Primary tests that have a short test administration time and a small number of test items. The reliability coefficients are comparable to those of other low-stakes language proficiency tests designed for young EAL learners such as TOEFL Junior (ETS, 2018b). The internal consistency reliability information for the TOEFL Primary Writing test was not available at the time of writing given the short tenure of the assessment. It will be provided when available.

 © 2024 Educational Testing Service

Research studies using the TOEFL Primary Reading and Listening tests as a measure of young EAL students' English language skills have also reported high internal consistency reliability. Chung and Révész (2024) used the entire reading section of the TOEFL Primary Reading and Listening Step 2 test to investigate the effect of textual enhancement of reading tasks on young EAL learners' grammatical development. Participants included 49 fourth graders from South Korea. The students' responses to the reading test items were scored dichotomously, giving one point for each correct answer and zero for an incorrect response. The researchers reported an alpha value of .90 based on the total number of correct answers or the raw scores. Hui et al. (2022), who examined the effect of reading-aloud input materials on children's listening test performance, reported internal consistency reliability of .86 based on 25 reading items and .82 based on 24 listening items. The reliabilities estimated in these two empirical studies exceeded thresholds traditionally considered acceptable (Knapp & Mueller, 2010), thus providing additional evidence supporting the assumption of adequately high internal consistency of the TOEFL Primary Reading and Listening tests.

Assumption 2

TOEFL Primary test scores are consistent across test administrations.

Backing. Analysis of repeat test-taker performance provides evidence supporting the stability of the TOEFL Primary test scores across test administrations.

In a study conducted by Cho and Blood (2020), empirical evidence was gathered regarding test-retest reliability of the TOEFL Primary Listening and Reading tests. The researchers analyzed score changes of 5,213 repeat test takers from Japan and Türkiye. These students had taken the tests two or three times, typically over a period of six months or less, although up to 12 months or more for some students in Japan. For the two-time repeaters, the analysis revealed minimal score changes. On average, the reading score point change was 0.65 for the Japanese and 0.70 for the Turkish students; the listening score point change was 1.02 for the Japanese and 0.85 for the Turkish students. These changes were smaller than the standard error of measurement (SEM) of each test section scale score (SEMs are 1.50 and 1.19 for TOEFL Primary Reading Step 1 and Step 2 tests, respectively; 1.36 and 1.22 for Listening Step 1 and Step 2 tests, respectively; ETS, 2019c, p. 29), supplying evidence that supports the stability of the test scores over repeated administrations.

 © 2024 Educational Testing Service

Assumption 3

Standard procedures are used to ensure score consistency across test forms.

Backing. Review of internal documents provides evidence that procedures for scaling and equating are well specified and adhered to.

This assumption is supported by a review of internal documentation related to ongoing scaling and equating of the TOEFL Primary Reading and Listening test forms (ETS, 2016a). The scaling and equating procedure is carried out using an IRT framework to estimate item parameters. For each test section, the total number of correct responses, the raw score, is statistically adjusted, or equated, to account for differences in difficulty between test forms. The equating process compensates for small differences across test forms and allows scores from each test form to be interpreted interchangeably. A team of psychometricians routinely conduct these procedures to ensure score consistency across different TOEFL Primary Reading and Listening test forms.

For the TOEFL Primary Speaking and Writing tests, the test results are not converted to scale scores. Instead, the speaking test result is the total of individual task scores assigned by human raters using standardized scoring rubrics (ETS, 2022a). The writing test result is the total of the individual task scores assigned by a machine. Score comparability of the Speaking and Writing tests is maintained through multiple measures, including small-scale item try-outs before operational use, rigorous rater training and certification (ETS 2018a, 2019b, 2023b), and monitoring the quality of automated scoring (Li & Futagi, 2023).

Assumption 4

Clearly defined item specifications are used to create parallel test items and tasks.

Backing. Item writers use well-defined item specifications to produce parallel test items and tasks.

The TOEFL Primary item writers adhere to well-defined item specifications for creating reading (ETS, 2022c), listening (ETS, 2022d), speaking (ETS, 2016b), and writing (ETS, 2023c) items. The specifications delineate the test constructs, task types, task characteristics, and test design principles as well as best practices in item authoring. The item specifications also outline which aspects of the test tasks must remain constant to consistently assess relevant language knowledge and skills, as well as aspects that may vary to create new items (e.g., the scenarios for the writing tasks). Example tasks are included in the item specifications as a reference.

 © 2024 Educational Testing Service

Explanation

The explanation inference is based on the warrant that TOEFL Primary test scores can be attributed to the constructs of social and academic ELP in primary and lower secondary education. Sources of backing for this warrant include investigations of the linguistic knowledge, processes, and strategies underlying test performance, the influence of task characteristics on test performance, and the relationship between test performance and experience in learning English.

Assumption 1

The strategies and processes required to successfully complete the TOEFL Primary tasks vary in keeping with theoretical expectations.

Backing. Task completion strategies and cognitive processes in which test takers engage are consistent with expected processes for successfully completing the TOEFL Primary tasks.

Backing for this assumption was found in studies that examined strategy use and cognitive processes in which test takers engaged when responding to the TOEFL Primary test tasks. Gu and So (2017) investigated strategy use during the prototyping stage to examine the extent to which the proposed reading and listening test items measured the intended constructs. Participants were 16 young EAL students from an after-school English language learning program in China. The students responded to sample prototype items and were interviewed about their strategy use during task performance. Drawing upon a theoretical framework of test-taking strategy by Cohen and Upton (2006), the researchers categorized the reported strategies into language-learner strategies, test-management strategies, and test-wiseness strategies. The findings indicated that the majority (63%) of the reported strategies belonged to the language-learner and test-management categories, relevant to the constructs measured.

Strategy use during speaking task performance was examined in a study conducted by Hsieh and Gu (2020). Verbal report data were gathered from 31 fourth to sixth grade students from a primary school in Hong Kong. The students were tasked with completing two TOEFL Primary and one TOEFL Junior picture-based speaking tasks. Upon completing each speaking task, they provided retrospective verbal insights into their strategy use during task execution. Analysis of the verbal report data revealed that the students predominately employed construct-relevant organizational and elaboration strategies. Furthermore, older students appeared to use more sophisticated strategies compared to their younger counterparts. The findings are consistent with existing theoretical frameworks regarding children's strategy use in picture-based narratives

 © 2024 Educational Testing Service

(Bamberg & Damrad-Frye, 1991), and provide empirical evidence of construct-relevant strategy use during speaking task performance.

Writing processes that young EAL learners engage in when responding to the computer-delivered TOEFL Primary Writing test tasks were explored in Wolf, Deane, et al. (2023). The researchers employed the keystroke logging methodology to capture students' keyboarding activities during writing task performance. Analysis of 1,335 young EAL learners' keystroke features demonstrated that the students engaged in construct-relevant writing behaviors, including prompt processing, planning, and editing, aligning with theoretical expectations of writing processes (Kellogg, 1996).

Assumption 2

Task difficulty is systematically influenced by task characteristics.

Backing. Empirical research findings suggested that the difficulty of TOEFL Primary test tasks was systematically influenced by task characteristics as expected.

Throughout test development, the TOEFL Primary test developers carefully examined the characteristics of the proposed test tasks and their implications for task difficulty. For example, Cho and So (2014) conducted a study involving 12 Korean EAL learners between 9 and 11 years of age to identify construct-irrelevant task characteristics that impacted task difficulty and test performance. The students responded to sample reading or listening prototype items and participated in retrospective interviews about their perceived task difficulty and cognitive processes during task execution. The students' verbal reports provided insights into factors that influenced perceived task difficulty, such as linguistic complexity of test questions, memory capacity of children, and wording of task instructions. Findings guided decisions regarding revisions to item specifications and task design, ensuring appropriate task difficulty and design for the target population.

Getman (2020) investigated the influence of task characteristics on the difficulty of the TOEFL Primary speaking pilot items. Analysis of test-taker performance showed that *Describe a Picture* task was the easiest whereas the *Give Directions* task was the most difficult. As expected, the speaking tasks that contained the element of topic choice were found to be easier than tasks that did not contain it.

Findings of the TOEFL Primary prototyping and pilot-testing studies provide empirical evidence demonstrating the systematic impact of task characteristics on task difficulty as

 © 2024 Educational Testing Service

predicted by theoretical models of language proficiency (Bachman & Palmer, 2010; Iwashita et al., 2001).

Assumption 3

Test performance varies according to experience in learning English.

Backing. Research results showed that test performance was related to the length of English instruction.

The length of time spent learning a language is an important factor affecting children's language acquisition (Collier, 2001) and TOEFL Primary test scores are expected to vary with the amount of language instruction that young EAL learners receive. Backing for this assumption first came from examination of repeated test-taker performances on the TOEFL Primary tests. Cho and Blood (2020) compared performances of more than 5,000 repeat test takers from Japan and Türkiye. Using the students' age as a proxy for years of English instruction given that the majority of the students started learning English at the same age or grade level, the results showed positive, high correlations between age and test scores. Older students, who were likely to have studied English for a longer period of time, obtained higher scores on the TOEFL Primary Reading and Listening tests compared to the younger ones. Though limited contextual information regarding test takers' actual learning experiences was available, the study provides some evidence suggesting that TOEFL Primary test scores reflect performance differences attributable to English instruction.

In another study, Wolf, Lopez, and Lee (2023) analyzed the repeated test performances of students in Grades 3 to 6 from five Turkish primary schools. The participating students took the TOEFL Primary Reading and Listening tests twice, first at the start of the school year and second toward the end. The average test scores in both reading and listening sections increased between the first and second administrations across all five schools, demonstrating the relationship between test performance and English language learning experience.

Further evidence was found in a joint research study between ETS and the Kenya Institute of Curriculum Development conducted by Hsieh et al. (2018). The researchers used the TOEFL Primary Reading and Listening Step 2 test to assess the ELP of Kenyan primary school students. The research team collected performance data from 4,768 Grades 3 to 7 students from 51 primary schools across Kenya. The students' average reading and listening test scores exhibited consistent improvement with each successive grade level. This upward trend in test

 © 2024 Educational Testing Service

performance was attributed to the cumulative effect of longer exposure to English language instruction within the Kenyan primary school context.

Extrapolation

The extrapolation inference is based on the warrant that the constructs of ELP as assessed by the TOEFL Primary tests reflect students' English language performance in primary and lower secondary school settings. Underlying this warrant is the assumption that test-taker performance is related to other criteria of language proficiency in the target domain of language use. Backing for this assumption requires evidence from criterion-related correlations between performances on the TOEFL Primary tests and other measures of ELP for young EAL learners.

Assumption

Performance on the TOEFL Primary tests is related to other criterion measures of ELP in the TLU domain.

Backing 1. Research showed that TOEFL Primary Reading test scores were related to an external measure of reading comprehension.

In a joint research effort between MetaMetrics, the creator of the Lexile Framework for Reading, and ETS (MetaMetrics, 2013), a team of researchers conducted a study to examine how well students' performances on the TOEFL Primary Reading test related to the Lexile Linking Test that measures children's reading comprehension skills such as identifying details in the passage, drawing conclusions, and making comparisons. A set of Lexile Linking Test items that had similar levels of item difficulty to those of the TOEFL Primary Reading test items was created for the study. These items and the TOEFL Primary Reading test were administered to over 2,200 young EAL students around the world. The correlation between test takers' performances on the two tests was .98, providing evidence showing that the TOEFL Primary Reading test was related to an external test of reading comprehension that measures similar constructs.

Backing 2. Research showed that performances on the TOEFL Primary tests were related to a sequence of language courses designed for young EAL learners.

Placement into different levels of language courses provides an indicator of students' language abilities in the classrooms (Chapelle et al., 2008). Thus course level can be used as an external criterion measure in support of the extrapolation assumption. Papageorgiou et al. (2020) gathered test performance data on the TOEFL Primary Reading and Listening tests from 5,440

 © 2024 Educational Testing Service

young EAL students to investigate the alignment between the TOEFL Primary tests and the content of an online-delivered language learning curriculum. The students were enrolled in the online course that spanned five different levels. Findings showed that the students' average reading and listening scores increased as the course level went up. Higher-performing students tended to enroll in the upper-level courses that emphasized more advanced language knowledge and skills and academic English. The result was consistent with expectations, supporting extrapolation of test scores to a non-test criterion.

Utilization

The utilization inference is based on the warrant that TOEFL Primary test scores are useful for measuring students' ELP, monitoring progress, and supporting teaching and learning. This warrant relies on the assumptions that the test scores are clearly interpretable by stakeholders and are appropriate for the intended uses. Evidence was collected from the availability of interpretive materials, stakeholder perceptions of the intended uses, and empirical studies of test-taker performances.

Assumption 1

The meaning of test scores is clearly interpretable by test takers, teachers, parents, and score users.

Backing. Interpretive materials are made available to help stakeholders understand score meanings and make decisions.

The TOEFL Primary testing program provides information about the test constructs, score meanings, sample score reports, sample test items, and proficiency descriptors on the TOEFL Primary website (https://www.ets.org/toefl/primary/about.html) and in test-taker handbooks (ETS, 2019c, 2023a). These interpretive materials are developed and reviewed by multiple groups, including ETS researchers, assessment developers, marketing, and business staff, prior to publication to ensure clarity, understandability, and relevance to the intended audience.

To enhance score interpretation and use, the TOEFL Primary test scores have been linked to the CEFR using standard-setting methodologies. The CEFR linking results provide teachers, institutions, and other score users a tool for interpreting TOEFL Primary test scores against a widely used international language proficiency framework and help stakeholders better understand score meanings and make score-based decisions for young EAL learners.

 © 2024 Educational Testing Service

Baron and Papageorgiou (2014) conducted a standard-setting study to link the scores of TOEFL Primary Reading and Listening and TOEFL Primary Speaking tests to levels of the CEFR. The research team first identified relevant CEFR level descriptors and revised them to reflect the performance of young EAL learners. The modified level descriptors, along with test-taker responses that provided direct evidence of test-takers' English language skills, were used as the linking materials for expert judgments. Eighteen language educators, experienced in teaching young EAL learners, from 15 countries served on the standard-setting panel and made recommended cut scores for reaching the A1, A2, B1, and B2 levels of the CEFR.

The TOEFL Primary Writing test scores were linked to levels of the CEFR in a separate standard-setting study (Suhan et al., 2024). A panel of 13 ETS assessment developers with backgrounds in English teaching and assessment served on the standard-setting panel. To ensure valid score mapping and interpretation the research team selected the CEFR level descriptors that were developed for young learners (Council of Europe, 2018a, 2018b) and relevant to the test constructs for expert judgments. As is with other TOEFL Primary tests, the panelists also made recommended cut scores for reaching the A1, A2, B1, and B2 levels of the CEFR.

For both standard-setting studies, the panelists were carefully recruited and trained prior to the standard-setting meetings to ensure proper representation and research rigor. Well-established procedures and standard-setting methods, including the modified Angoff procedure (Cizek & Bunch, 2007) and Performance Profile method (Fleckenstein et al., 2020; Hambleton et al., 2000), were systematically followed during the standard-setting meetings. The project teams also collected evidence addressing the validity of the standard-setting processes by gathering panelists' perceptions on their own standard-setting judgments. Overall, high-level of procedural validity evidence was obtained and reported.

Assumption 2

TOEFL Primary test scores are useful for the intended purposes.

Backing 1. Stakeholders perceived the TOEFL Primary tests as useful for the intended purposes.

Teachers and students are key stakeholders of the TOEFL Primary tests and their perceptions provide valuable insights into the usefulness of the intended purposes. Hsieh and Ionescu (2015) sought input from primary school teachers in Kenya regarding the usefulness of the TOEFL Primary Reading and Listening tests. The participants were 17 classroom teachers

 © 2024 Educational Testing Service

from six geographical locations in Kenya. The teachers reviewed test description materials and sample items and participated in focus group or one-on-one interviews. All the teachers viewed the tests as well-designed with appropriate progression of text complexity and difficulty for young EAL learners and thought that the tests would be a valuable tool for assessing their students' language proficiency. The teachers also held positive views about the feedback provided in the score report (i.e., stars, badges) and emphasized its usefulness for providing direction and motivation for further learning.

Stakeholder perception of test usefulness was also explored in Wolf, Lopez, and Lee (2023). The researchers interviewed classroom teachers multiple times over a two-year period, and gathered survey responses from teachers, parents, and students to explore their perspectives regarding the use of the tests to measure students' ELP, monitor progress, and support teaching and learning. The participating teachers and parents viewed the TOEFL Primary tests as particularly useful for progress monitoring. The students were positive about the use of the tests to measure their language ability and thought that the tests provided useful feedback and boosted their learning motivation.

Backing 2. TOEFL Primary tests have positive impact on teaching and learning.

To support teaching and learning, the TOEFL Primary testing program offers resources and tools to assist with teaching and help students in their English language learning. A book of worksheets that contains independent and interactive practice activities in reading, listening, and speaking skills assessed on the TOEFL Primary tests are made freely available to language educators (ETS, 2023d). In addition, ETS offers an online learning program, the TOEFL Primary English Learning Center (ELC), with interactive activities for skill building and automated feedback on listening and reading exercises. Teachers can monitor student progress and provide feedback through a teacher management system.

To gauge the effectiveness of the TOEFL Primary ELC, Timpe-Laughlin (2015) conducted an evaluation study with stakeholders in Brazil and Mexico. Data were collected from two curriculum coordinators, ten teachers, and 40 students. Findings revealed that educators considered the ELC a valuable test-preparation tool, beneficial for students' English language development. Additionally, students viewed the ELC activities favorably, finding them enjoyable and helpful for language learning.

 © 2024 Educational Testing Service

To investigate the washback of the TOEFL Primary tests on teaching and learning, Gu et al. (2016) conducted a longitudinal study with stakeholders at a private language institution in Peru over a two-year period. The researchers interviewed and surveyed four school administrators and nine classroom teachers twice during the study period. The teachers reported that they used the TOEFL Primary Reading and Listening tests primarily to assess students' language skills and provide parents with an international exam certificate. While the researchers found minimal direct impact on classroom instruction as a result of test use, the teachers expressed a positive perspective regarding the tests' contribution to enhancing students' motivation toward learning English.

In another washback study, Wolf, Lopez, and Lee (2023) employed Bailey's washback model (Bailey, 1999) as a systematic framework for data collection and analysis. The study results suggested that the teachers used the TOEFL Primary tests as an objective, external measure to evaluate students' progress in relation to the CEFR. Similar to Gu et al. (2016), limited washback on daily classroom practices, such as changes to teaching content or methodologies as a result of test utilization, was observed. Instead, macro-level washback effects, such as the provision of additional courses to support students in need and leveraging test results to inform class placement decisions, were found. These findings provide some support for the use of TOEFL Primary tests to inform language teaching and learning.

Discussion

This paper presents the TOEFL Primary validity argument by providing evidence and research backing for each of the validity inferences and assumptions proposed in the TOEFL Primary interpretive argument. Table 2 presents the validity argument with references to the backing. The collected evidence supports claims regarding the validity of the TOEFL Primary tests for assessing students' ELP monitoring progress, and supporting teaching and learning. This section summarizes and evaluates the main findings in relation to the inferences in the validity argument.

The domain description inference was supported through a series of domain analysis activities conducted during early stages of test development. These activities informed the identification of important TLU tasks, and the language knowledge and skills required to complete these tasks. Empirical investigations of teachers' perceptions of task relevance and

 © 2024 Educational Testing Service

importance serve as crucial evidence supporting the claim that the assessment tasks are relevant and representative of the instructional domain targeted by the TOEFL Primary tests.

Table 2. TOEFL Primary Validity Argument

Inference	Warrants	Assumptions	Sources of backing
Domain description	TOEFL Primary test tasks represent relevant language knowledge and skills in situations representative of those in the TLU domain.	1. Assessment items and tasks produce evidence of language knowledge and skills representative of those used in the TLU domain.	Chen & Sheehan (2015) Cho et al. (2016) Hsieh et al. (2013) Papageorgiou et al. (2020) Turkan & Alder (2011) Wolf et al. (2024)
		2. Teachers perceive the test tasks as relevant and representative of those taught in the TLU domain.	Gu et al. (2016) Hsieh (2013, 2016)
Evaluation	Scores on the TOEFL Primary tests reflect the targeted language knowledge and skills.	1. Scoring procedures and guidelines are well-specified and followed.	ETS (2016a, 2018a, 2019b, 2023b)
		2. Scoring rubrics are appropriate for judging evidence of the targeted language knowledge and skills.	Getman (2020) Li & Futagi (2023) Wolf et al. (2024) Wolf & Suhan (2023)
		3. Statistical characteristics of test items and tasks are appropriate for providing evidence of the targeted language knowledge and skills.	Cho et al. (2016, 2017) ETS (2022b) Getman (2020, 2024) Zu et al. (2017)
		4. Task administration conditions are appropriate for providing evidence of the targeted language knowledge and skills.	Ballard (2016) Cho et al. (2016) Getman et at. (2016) Hui et al. (2022) Lee & Winke (2018) Wolf et al. (2024)
Generalization	Similar scores on the TOEFL Primary tests are expected to be obtained across comparable tasks, test forms, administrations,	1. Sufficient numbers of items and tasks are included to provide reliable estimates of test-taker performance.	Cho et al. (2016, 2017) Chung & Révész (2024) ETS (2018, 2019b) Getman (2020, 2024) Hui et al. (2022) Zu et al. (2017)
		2. TOEFL Primary scores are consistent across test administrations.	Cho & Blood (2020)

 © 2024 Educational Testing Service

	and testing conditions.	3.	Standard procedures are used to ensure score consistency across test forms.	ETS (2016a, 2018a, 2019b, 2022a, 2023b) Li & Futagi (2023)
		4.	Clearly defined item specifications are used to create parallel items and tasks.	ETS (2016b, 2022c, 2022d, 2023c)
Explanation	TOEFL Primary scores can be attributed to the constructs of social and academic English language ELP in primary and lower secondary-level education.	1.	The strategies and processes required to successfully complete the TOEFL Primary tasks vary in keeping with theoretical expectations.	Gu & So (2017) Hsieh & Gu (2020) Wolf, Deane, et al. (2023)
		2.	Task difficulty is systematically influenced by task characteristics.	Cho et al. (2016) Cho & So (2014) Getman (2020)
		3.	Test performance varies according to experience in learning English.	Cho & Blood (2020) Hsieh et al. (2018) Wolf, Lopez, and Lee (2023)
Extrapolation	TOEFL Primary test scores reflect students' performances in the TLU domain in primary and lower secondary school contexts.	1.	Performance on the TOEFL Primary tests is related to other criterion measures of ELP in the TLU domain.	MetaMetrics (2013) Papageorgiou et al. (2020)
Utilization	TOEFL Primary tests are useful for measuring students' language proficiency, monitoring progress, and supporting teaching and learning.	1.	The meaning of test scores is clearly interpretable by test takers, teachers, parents, and other score users.	Baron & Papageorgiou (2014) ETS (2019c, 2023a) Suhan et al. (2024)
		2.	TOEFL Primary test scores are appropriate for the intended purposes.	Cho & Blood (2020) ETS (2023d) Gu et al. (2016) Hsieh & Ionescu (2015) Timpe-Laughlin (2015) Wolf, Deane, et al. (2023) Wolf, Lopez, and Lee (2023)

Note. ELP = English language proficiency; TLU = target language use.

 © 2024 Educational Testing Service

It is worth pointing out that due to the rapid change and advances in technology, the way with which young EAL learners learn and use English language for social and academic purposes in the target domain of language use may have evolved in recent years. For example, children may use mobile language applications to interact with learning partners or AI agents and receive immediate feedback in and outside EAL classrooms, anytime, anywhere. With the changing nature of language learning landscape, the application of technology to assess the language skills of and provide feedback to young EAL learners is an avenue of research needed in future domain analysis effort to ensure that the TOEFL Primary assessment tasks continue to reflect the important KSAs in the TLU domain. Another area that requires further investigation pertains to the extent to which technology-mediated language learning and assessment has shifted stakeholder perceptions of the relevance and importance of the TOEFL Primary assessment tasks and the implications for score interpretation and use.

For the evaluation inference, evidence was obtained to demonstrate that the TOEFL Primary test scores reflect the targeted language abilities and that the test items and scoring procedures are adequate for the intended interpretations. This claim was supported by information outlined in internal documentation pertaining to standard scoring procedures, statistical analyses of test items, rater training and certification, as well as empirical investigations into tasks, scores, and various test administration conditions. Findings from these documents and empirical studies provide backing for the assumption that TOEFL Primary test scores accurately assess the targeted language skills and knowledge. Nonetheless, it should be pointed out that one study found that the countdown clock presented in the TOEFL Primary Speaking test tasks could impact test-taker performance (Lee & Winke, 2018). Future research is required to examine the impact of the countdown clock on students' *actual* test scores under operational conditions to ensure appropriate speaking task administration conditions.

The generalization inference necessitates evidence showing that test takers would consistently receive similar test scores across comparable assessment tasks, test forms, test administrations, and scoring conditions. Backing for this claim was obtained from a range of sources, including findings from reliability studies, standardized scaling and equating procedures, and the use of item specifications to construct parallel test items and tasks. This collection of validity evidence supports assertions regarding the consistency of test scores and their interpretations. Of note is that further research is critical to assess the efficacy of utilizing

 © 2024 Educational Testing Service

the AWE engine to score the TOEFL Primary Writing test responses, given that the extent to which writing abilities can be validly scored by machines remains under-researched in the assessment of young EAL learners (Hannah et al., 2023).

The explanation inference links the test items and scores back to the test construct definition. To assess the alignment with theoretical expectations of performance on the TOEFL Primary tests, evidence was gathered through research on response analysis, cognitive processes and strategic behaviors, task characteristics and their impact on task difficulty, as well as the connection between test performance and learning experience. The amassed empirical evidence supports the assumption that the scores yielded by the TOEFL Primary tests are attributable to the theoretical constructs targeted by the tests.

The extrapolation inference extends the connection between test scores and test-taker performance into the TLU domain. Empirical evidence is presented to provide links between the TOEFL Primary test scores and external criterion measures of language proficiency, including a reading comprehension test and levels of an EAL course. Despite the evidence, more empirical research in this area is needed to provide stronger backing for the assumption that performance on the TOEFL Primary tests reflects what young EAL learners can do with English in the TLU domain. To this end, additional criterion-related validity evidence can be gathered from connecting test performance to teachers' judgements of students' abilities, students' self-assessments, and classroom-based performances, among others. Collecting data, such as self-assessment, with school-aged EAL learners can be challenging; care should be taken to ensure that valid data with young learners can be gathered and that confounding effects such as young learners' individual differences in cognitive development are considered in the research design.

The utilization inference concerns the usefulness and appropriateness of the TOEFL Primary tests for the intended score interpretations and uses. The test scores are conveyed to stakeholders through various channels including score reports, test-taker handbooks, websites, and research documents to ensure accurate interpretations. The cumulative evidence affirms that TOEFL Primary tests are useful for assessing students' ELP, monitoring progress, and supporting teaching and learning. Nonetheless, limited direct evidence was found with regard to positive impact on daily language instruction. Future research will be needed to understand how classroom teachers can more effectively utilize the tests to make instructional decisions based on students' test performance. To further support teaching and learning, test developers may

 © 2024 Educational Testing Service

consider providing additional feedback or more fine-grained information about students' performances on different language subskills (e.g., understanding simple written narratives), although such information should be offered with sound psychometric justifications for providing subscores (Choi & Papageorgiou, 2020; Papageorgiou & Choi, 2018). It is equally important to investigate whether providing subscores leads to a better understanding and more appropriate use of the test results by teachers and other stakeholders.

Finally, it should be emphasized that additional validation research on the newly launched, computer-mediated TOEFL Primary Writing test is required to support score use and interpretation. The test design framework paper (Wolf et al., 2024) proposes a potential list of research topics that need validity research. Adding to the topics proposed, as AI capabilities continue to improve, exploring the use of AWE systems to provide more targeted, learning-oriented feedback or insights that can inform teaching and learning is another fruitful path for validation research.

Conclusion

This paper presents a fully articulated TOEFL Primary interpretive argument as a roadmap for the collection of evidence and evaluation of the TOEFL Primary validity argument. A considerable body of evidence was found to back up the claims that TOEFL Primary tests are appropriate and useful for measuring students' ELP, monitoring progress, and supporting teaching and learning. However, there were some limitations with respect to the task administration conditions and washback effects—areas that need to be addressed in future research.

The TOEFL Primary validity argument contributes to the demand for validation of standardized language proficiency tests designed for young EAL students using an argument-based approach to validation (Fan & Yan, 2020). It serves as an example to illustrate how the argument-based framework can be applied to validate low-stakes young EAL language assessments. The TOEFL Primary validity argument also provides researchers and stakeholders with information about the test quality and can guide them to make decisions about the intended interpretations and uses of the test scores and their suitability for different learning contexts.

 © 2024 Educational Testing Service

References

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Bailey, K. M. (1999). *Washback in language testing* (Research Memorandum No. RM-99-04). ETS. https://www.ets.org/Media/Research/pdf/RM-99-04.pdf

Ballard, L. (2016). *Young ELLs and computerized reading assessment: Are kids ready?* [Paper presentation]. Midwest Association of Language Testers (MwALT) Conference, West Lafayette, Indiana.

Bamberg, M., & Damrad-Frye, R. (1991). On the ability to provide evaluative comments: Further explorations of children's narrative competencies. *Journal of Child Language, 18*(3), 689–710. https://doi.org/10.1017/s0305000900011314

Baron, P., & Papageorgiou, S. (2014). *Mapping the TOEFL® Primary™ Test onto the Common European Framework of Reference* (Research Memorandum No. RM-14-05). ETS. https://www.ets.org/Media/Research/pdf/RM-14-05.pdf

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Chen, J., & Sheehan, K. M. (2015). *Analyzing and comparing reading stimulus materials across the TOEFL® family of assessments* (Research Report No. RR-15-08). ETS. https://doi.org/10.1002/ets2.12055

Cho, Y., & Blood, I. (2020). An analysis of TOEFL Primary® repeaters: How much score change occurs? *Language Testing, 37*(4), 503–522. https://doi.org/10.1177/0265532220927751

Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2016). *Designing the TOEFL® Primary™ Tests* (Research Memorandum No. RM-16-12). ETS. https://www.ets.org/Media/Research/pdf/RM-16-12.pdf

Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2017). Designing the TOEFL Primary tests. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 41–58). Routledge. https://doi.org/10.4324/9781315674391-3

Cho, Y., & So, Y. (2014). *Construct-irrelevant factors influencing young English as a foreign langauge (EFL) learners' perceptions of test task difficulty* (Research Memorandum No. RM-14-04). ETS. https://www.ets.org/Media/Research/pdf/RM-14-04.pdf

 © 2024 Educational Testing Service

Choi, I., & Papageorgiou, S. (2020). Evaluating subscore uses across multiple levels: A case of reading and listening subscores for young EFL learners. *Language Testing, 37*(2), 254–279. https://doi.org/10.1177/0265532219879654

Chung, Y., & Révész, A. (2024). Investigating the effect of textual enhancement in post-reading tasks on grammatical development by child language learners. *Language Teaching Research*, *28*(2), 632–653. https://doi.org/10.1177/13621688211005068

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Sage. https://doi.org/10.4135/9781412985918

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the new TOEFL reading tasks* (Research Report No. RR-06-06). ETS. https://doi.org/10.1002/j.2333-8504.2006.tb02012.x

Collier, V. P. (2001). How long? A synthesis of research on academic achievement in a second langauge. In M. M. Suárez-Orozco, C. Suárez-Orozco, & D. Qin-Hilliard (Eds.), *Interdisciplinary Perspectives on the New Immigration Volume 6: The new immigrant and language* (pp. 49–71). Routledge. https://doi.org/10.4324/9781315054223

Common Core State Standards Initiative (2010). *Common core state standards for English language arts & literacy in history/social studies, science and technical subjects*. CCSSO & National Governors Association.

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2018a). *Collated representative samples of descriptors of language competences developed for young learners. Volume 1: Ages 7–10.* https://rm.coe.int/16808b1688

Council of Europe. (2018b). *Collated representative samples of descriptors of language competences developed for young learners. Volume 2: Ages 11–15.* https://rm.coe.int/16808b1689

ETS. (2014). *ETS standards for quality and fairness*. https://www.ets.org/pdfs/about/standards-quality-fairness.pdf

ETS. (2016a). *TOEFL Primary Reading and Listening statistical procedures for psychometricians* [Unpublished internal company document].

ETS. (2016b). *TOEFL Primary Speaking Item Specifications (Updated)* [Unpublished internal company document].

ETS. (2018a). *Online Network for Evaluation (ONE): ONE training* [Unpublished internal company document].

ETS. (2018b). *Handbook for the TOEFL Junior® tests*. https://www.ets.org/pdfs/toefl/toefl-junior-handbook.pdf

ETS. (2019a). *TOEFL Primary® framework and test development*. TOEFL Research Insight Series, Volume 8. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v8.pdf

ETS. (2019b). *Online Network for Evaluation (ONE): Leader monitoring* [Unpublished internal company document].

ETS. (2019c). *Handbook for the TOEFL Primary® tests*.

ETS. (2022a). *TOEFL Primary Speaking Scoring Guide.* https://www.ets.org/s/toefl_primary/pdf/toefl_primary_speaking_scoring_guide.pdf

ETS. (2022b). *TOEFL Primary® and TOEFL Junior® Tests score data summary: For 2016-2018*. https://www.ets.org/pdfs/toefl-junior-score-data-summary.pdf

ETS. (2022c). *TOEFL Primary Reading item specifications (Updated)* [Unpublished internal company document].

ETS. (2022d). *TOEFL Primary Listening item specifications (Updated)*. [Unpublished internal company document].

ETS. (2023a). *TOEFL Primary® test taker handbook*. https://www.ets.org/pdfs/toefl/toefl-primary-handbook.pdf

ETS. (2023b). *Rater training manual for the TOEFL Primary® Speaking test*. [Unpublished internal company document].

ETS. (2023c). *TOEL Primary Writing item specifications*. [Unpublished internal company document].

ETS. (2023d). *TOEFL Primary Activity Worksheets*.
https://www.ets.org/content/dam/ets-org/pdfs/toefl/toefl-ibt-insight-s1v11.pdf

ETS. (2024). *TOEFL Primary*. https://www.ets.org/toefl/primary/about.html

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.00330

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing, 43*, 100420. https://doi.org/10.1016/j.asw.2019.100420

Getman, E. (2020). *Age, task characteristics, and acoustic indicators of engagement: Investigations into the validity of a technology-enhanced speaking test for young language learners* [Doctoral dissertation, Columbia University]. Columbia Academic Commons Repository. https://doi.org/10.7916/d8-ad10-bc50

Getman, E. (2024). Speaking test tasks for young language learners. In L. Davis & J. Norris, (Eds.), *Challenges and innovations in speaking assessment* (pp. 85–105). Routledge. https://doi.org/10.4324/9781315201504-7

Getman, E., Cho, Y., & Luce, C. (2016). *Effects of printed option sets on listening item performance among young English-as-a-foreign-language learners* (Research Memorandum No. RM-16-16). ETS. https://www.ets.org/Media/Research/pdf/RM-16-16.pdf

Gu, L., Hsieh, C.-N., & Tolentino, F. (2016). *Examining the impact of the TOEFL Primary tests on teaching and learning for young EFL learners* [Unpublished manuscript]. ETS.

Gu, L., & So, Y. (2017). Strategies used by young English learners in an assessment context. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 118–135). Routledge. https://doi.org/10.4324/9781315674391-7

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355–366. https://doi.org/10.1177/01466210022031804

Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). Validity arguments for automated essay scoring of young students' writing traits. *Language Assessment Quarterly, 20*(4), 5. https://doi.org/10.1080/15434303.2023.2288253

Hsieh, C.-N. (2013). *Establishing domain representations for a large-scale language assessment for young EFL learners* [Paper presentaiton]. Midwest Association of Language Tester (MwALT) Conference, East Lansing, Michigan.

Hsieh, C.-N. (2016). Examining content representativeness of a young learner language assessment: EFL teachers' perspectives. In M. Nikolov (Ed.), *Assessing young learners*

of English: Global and local perspectives (pp. 262–284). Springer. https://doi.org/10.1007/978-3-319-22422-0_5

Hsieh, C.-N. (2024). *Building a validity argument for the TOEFL Junior® tests* (Research Report No. RR-24-05). ETS. http://doi.org/10.1002 /ets2.12379

Hsieh, C.-N., Cho, Y., Luce, R., & Wang, Y. (2013). *Alignment of the Sesame English Street program with TOEFL Young Learners assessment products* [Unpublished manuscript]. ETS.

Hsieh, C.-N., & Gu, L. (2020). Young language learners' strategy use and perceptions of picture-based speaking tasks. In R. Damerow & K. M. Bailey (Eds.), *Chinese-speaking learners of English: Research, theory, and practice* (pp. 171–182). Routledge. https://doi.org/10.4324/9780429290848-14

Hsieh, C.-N., & Ionescu, M. (2015). *Examining the use of TOEFL Primary for young English learners in the Kenyan context* [Unpublished manuscript]. ETS.

Hsieh, C.-N., Ionescu, M., & Ho, T.-H. (2018). Out of many, one: Challenges in teaching multilingual Kenyan primary students in English. *Language, Culture and Curriculum, 31*(2), 199–213. https://doi.org/10.1080/07908318.2017.1378670

Hui, B., Wong, S. S. Y., & Au, R. K. C. (2022). Reading aloud listening test items to young learners: Attention, item understanding, and test performance. *System, 108,* 102831. https://doi.org/10.1016/j.system.2022.102831

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401–436. https://doi.org/10.1111/0023-8333.00160

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.

Kane, M. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. (2021). Articulating a validity argument. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 34–47). Routledge. https://doi.org/10.4324/9781003220756-4

Kellogg, R. T. (1996). A model of working memory in writing. In M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–72). Lawrence Erlbaum Associates.

Knapp, T. R., & Mueller, R. O. (2010). Reliability and validity of instruments. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 337–341). Routledge.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing, 35*(2), 239–269. https://doi.org/10.1177/0265532217704009

Leung, C. (2022). English as an additional language: A close-to-practice view of teacher professional knowledge and professionalism. *Language and Education, 36*(2), 170–187. https://doi.org/10.1080/09500782.2021.1980003

Li, F., & Futagi, Y. (2023). *Evaluating the automated scoring of the TOEFL Junior® and TOEFL Primary® writing tests* [Paper presentation]. 2023 TOEFL® Research Symposium, Princeton, NJ.

MetaMetrics. (2013*). Linking the TOEFL Primary with the Lexile® Framework: A study to link the TOEFL® Primary™ Test with the Lexile® Framework for Reading* [Unpublished manuscript]. MetaMetrics.

MetaMetrics. (2024). *Lexile Framework for Reading*. https://lexile.com/

Nguyen, T. M. H., Gu, P., & Coxhead, A. (2024). Argument-based validation of academic collocation tests. *Language Testing*, *41*(3), 459–505. https://doi.org/10.1177/02655322231198499

Papageorgiou, S., & Choi, I. (2018). Adding value to second-language listening and reading subscores: Using a score augmentation approcah. *International Journal of Testing, 18*(3), 207–230. https://doi.org/10.1080/15305058.2017.1407766

Papageorgiou, S., Xu, X., Timpe-Laughlin, V., & Dugdale, D. M. (2020). *Exploring the alignment between a curriculum and a test for young learners of English as a foreign*

language (Research Memorandum No. RM-20-08). ETS. https://www.ets.org/Media/Research/pdf/RM-20-08.pdf

Schmidgall, J. (2017). *Articulating and evaluating validity arguments for the TOEIC® tests* (Research Report No. RR-17-51). ETS. https://doi.org/10.1002/ets2.12182

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards* (Research Report No. RR-10-28). ETS. https://doi.org/10.1002/j.2333-8504.2010.tb02235.x

Suhan, M., Papageorgiou, S., & Wolf, M. K. (2024). *Mapping the scores of the TOEFL Primary® Writing test to the Common European Framework of Reference levels* (Research Memorandum No. RM-24-03). ETS. https://www.ets.org/Media/Research/pdf/RM-24-03.pdf

Timpe-Laughlin, V. (2015). *Evaluating a learning tool for young English learners: The case of the TOEFL® Primary™ English Learning Center* (Research Memorandum No. RM-15-04). ETS. https://www.ets.org/Media/Research/pdf/RM-15-04.pdf

Toulmin, S. (2003). *The uses of argument* (2nd ed.). Cambridge. https://doi.org/10.1017/CBO9780511840005

Turkan, S., & Aldler, R. (2011). *Conceptual framework for the assessment of young learners of English as a foreign language* [Unpublished manuscript]. ETS.

Wolf, M. K., & Butler, Y. G. (2017). An overview of English language proficiency assessments for young learners. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 3–21). Routledge. https://doi.org/10.4324/9781315674391-1

Wolf, M. K., Deane, P., Li, C., Choi, I., & Suhan, M. (2023). *Young EFL students' writing processes during a computer-based assessment: An examination of keystroke logs* [Paper presentation]. Language Testing Research Colloquium (LTRC), New York.

Wolf, M. K., Lopez, A., & Lee, J. (2023). An investigation of the use of standardized and local assessments for young EAL students. In G. Brooks, J. Clenton, & S. Fraser (Eds.), *EAL research for the classroom: Practical and pedagogical implications* (pp. 164–184). Routledge. https://doi.org/10.4324/9781003274889

Wolf, M. I., & Suhan, M. (2023). *Providing scaffolding in a writing assessment for young EFL students* [Paper presentation]. TESOL Conference, Portland, Oregon.

Wolf, M. K., Suhan, M., Ginsburgh, M., Futagi, Y., & Li, F. (2024). *Design framework for the TOEFL Primary® writing test* (Research Memorandum No. RM-24-02). ETS. https://www.ets.org/Media/Research/pdf/RM-24-02.pdf

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199–225. https://doi.org/10.1177/0265532214557113

Zu, J., Moulder, B., & Morgan, R. (2017). A field test study for the TOEFL® Primary™. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 99–117). Routledge. https://doi.org/10.4324/9781315674391-6

Suggested citation:
Hsieh, C.-N. (2024). *Building a validity argument for the TOEFL Primary® tests* (TOEFL Research Report No. RR-103). ETS.

Action Editor: Larry Davis
Reviewers: Veronika Timpe-Laughlin and Mikyung Wolf

ETS, the ETS logo, TEXTEVALUATOR, TOEFL, TOEFL IBT, TOEFL JUNIOR, and TOEFL PRIMARY are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database.

www.ets.org

Copyright © 2024 by ETS. The Eight-Point logo is a trademark of ETS. All other trademarks are property of their respective owners.