ETS TOEFL

Quality Beyond Measure.

TOEFL® Research Report
TOEFL–RR-102
ETS Research Report No. RR–24-05

Building a Validity Argument for the TOEFL Junior® Tests

Ching-Ni Hsieh

December 2024

The *TOEFL*® test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT*® test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL*® *Primary*™ and *TOEFL Junior*® tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP*® Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL Family of Assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org **Web site:** www.ets.org/toefl

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

ETS RESEARCH REPORT

Building a Validity Argument for the TOEFL Junior® Tests

Ching-Ni Hsieh

ETS, Princeton, New Jersey United States

The TOEFL Junior® tests are designed to evaluate young language students' English reading, listening, speaking, and writing skills in an English-medium secondary instructional context. This paper articulates a validity argument constructed to support the use and interpretation of the TOEFL Junior test scores for the purpose of placement, progress monitoring, and evaluation of a test taker's English skills. The validity argument is built within an argument-based approach to validation and consists of six validity inferences that provide a coherent narrative about the measurement quality and intended uses of the TOEFL Junior test scores. Each validity inference is underpinned by specific assumptions and corresponding evidential support. The claims and supporting evidence presented in the validity argument demonstrate how the TOEFL Junior research program takes a rigorous approach to supporting the uses of the tests. The compilation of validity evidence serves as a resource for score users and stakeholders, guiding them to make informed decisions regarding the use and interpretation of TOEFL Junior test scores within their educational contexts.

Keywords TOEFL Junior® test; reading; listening; speaking; writing; validity; test scores; score users; English language assessment; young language students; English as a second language; English as a foreign language

doi:10.1002/ets2.12379

To meet the growing demand for assessing English as a second or foreign language (ESL/EFL) for young language learners (YLLs) worldwide, ETS developed and launched the TOEFL Junior® tests in 2010. The tests are designed for students ages 11 and older whose first language is not English and who are in the process of developing the English language proficiency (ELP) required to participate in an English-medium secondary instructional context, (i.e., the target language use [TLU] domain). The TOEFL Junior tests measure students' ability to use English for communicative purposes in situations and tasks representative of English-medium school contexts. The test scores are intended to provide information about the academic and social ELP of test takers to support decisions regarding placing students into different instructional levels and to monitor student progress in developing ELP over time. The tests can inform instruction in English-language programs that prepare students for academic English skills.

The TOEFL Junior TLU domain includes three subdomains — social-interpersonal, navigational, and academic — with more emphasis placed on the academic domain. These subdomains are informed by Bailey and colleagues' extensive research on school language (Bailey, 2007; Bailey & Heritage, 2008) and the TOEFL Junior test design team's review of language learning standards and curricula. The social-interpersonal subdomain encompasses uses of language for establishing and maintaining personal relationships, where, for example, students participate in casual conversations with friends or classmates in school settings. The navigational subdomain encompasses uses of language to communicate or navigate school or course information, where, for example, students communicate with peers or school staff about school- and course-related materials and activities but not about academic content. The academic subdomain entails language activities performed to learn academic content in English, such as participating in short conversations about academic content in a class, comprehending written academic texts, and summarizing oral or written academic texts (see So et al., 2015, for details).

The TOEFL Junior test constructs and task designs are guided by Bachman and Palmer's (2010) theoretical model of language knowledge. The model informed the test construct definition and the identification of important language knowledge and skills to measure, the features of reading and listening passages, and the expected characteristics of spoken and written performance (So et al., 2015). The theoretical model also illuminated the design and operationalization of the test tasks and allowed the test developers to effectively simulate the actual TLU tasks that would provide evidence about a test taker's language ability to communicate in English in the three TLU subdomains.

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

Since the initial launch, the TOEFL Junior tests have undergone several changes. Originally, two versions of the TOEFL Junior tests were developed to meet different test takers' and score users' needs, which included the paper-based and computer-delivered TOEFL Junior Standard test and the computer-delivered TOEFL Junior Comprehensive test. The two tests targeted the same TLU domain and subdomains described above, although they differed in the skills tested (see details in the next paragraph). Depending on stakeholders' needs, either the TOEFL Junior Standard or the TOEFL Junior Comprehensive test could be used. The TOEFL Junior Comprehensive test was sunset in 2016 due to operational considerations, though the speaking section of the test — TOEFL Junior Speaking — continued to be offered as a stand-alone test. In August 2023, the TOEFL Junior testing program launched a new, stand-alone test — TOEFL Junior Writing, which was largely developed based on the writing section of the TOEFL Junior Comprehensive test. Thus, at the time of writing, the TOEFL Junior tests include three separate tests: the paper-based and computer-delivered TOEFL Junior Standard, the computer-delivered TOEFL Junior Speaking, and the computer- delivered TOEFL Junior Writing (see https://www.ets.org/toefl/junior.html).

The paper-based and computer-delivered TOEFL Junior Standard test consists of three sections — Listening Comprehension, Language Form and Meaning, and Reading Comprehension. The Listening Comprehension section measures a test taker's ability to listen to and understand English. The Language Form and Meaning section measures a test taker's ability to demonstrate proficiency in key enabling English skills such as grammar and vocabulary in context. The Reading Comprehension section measures a test taker's ability to read and understand academic and non-academic texts written in English. Test questions in all three sections are multiple-choice questions. Test takers mark their responses to the test questions on an answer sheet. The answer sheet is then read by a machine for scoring. Each section contains 42 four-choice questions with a total testing time of 1 hour and 55 minutes. Each section score is determined by the number of questions a student has answered correctly. The number of correct responses on each section (i.e., the raw score) is converted to a scale score of 200–300 points. The total scale score is the sum of the three section scale scores and ranges from 600–900 points. The score report provides information about the total scale score, an overall score level accompanied by an overall performance descriptor, a description of the English-language abilities typical of test takers scoring within a particular scale score range, section scale scores and their corresponding levels on the Common European Framework of Reference (CEFR; Council of Europe, 2001) that shows the test taker's ability in comparison to a widely used tool for describing language proficiency, and a Lexile measure (see https://lexile.com/) based on the Reading Comprehension section scale score.

The TOEFL Junior Speaking test measures a test taker's ability to communicate orally in English in a classroom setting in secondary level education. The test is delivered on the computer and has four tasks and lasts about 18 minutes. The four tasks include one Read Aloud, one Picture Narration, one nonacademic Listen-Speak, and one academic Listen-Speak task. The responses are scored by trained human raters using a 0–4 point scoring rubric (ETS, 2022a); total score is reported on a scale of 0–16. The score report provides information about the total score (i.e., sum of the four task raw scores), performance descriptors of a test taker's ability, and the corresponding CEFR level. Detailed information about the speaking score level descriptors can be found at https://www.ets.org/pdfs/toefl/toefl-junior-speaking-descriptors.pdf.

The TOEFL Junior Writing test measures a test taker's computer-based English writing ability to communicate for social-interpersonal, school navigational, and academic purposes. The test has four task types, including Edit, E-mail, Opinion, and Listen-Write, and lasts about 40 minutes. The Edit task contains two sets of multiple-choice questions, and each of the other three tasks requires a written response. The written responses are scored by an automated scoring engine using natural language processing and artificial intelligence (AI) capabilities, which is trained on human ratings (ETS, 2023). As with Speaking, the total score ranges from 0 to 16 (See ETS, 2022b, for the writing scoring guide). The score report provides information about the total score and the accompanied overall performance descriptor, a description of the English-language abilities typical of test takers scoring within a particular score range, and the corresponding CEFR level. Detailed information about the writing score level descriptors can be found at https://www.ets.org/pdfs/toefl/toefl-junior-writing-score-descriptors.pdf.

Argument-Based Approach to Validation

The TOEFL Junior validity argument is built using an argument-based approach to validation (Kane, 2006, 2013) and draws on the TOEFL iBT® validity argument detailed in Chapelle et al. (2008). Kane (1992) first introduced the argument-based validity framework, using Toulmin's (1958, 2003) framework of argumentation, where claims about a test are supported by warrants (i.e., general statements that are used to justify a claim) or contradicted by rebuttals

(i.e., 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. The framework was later expanded in Kane (2006, 2013) where he identified two types of arguments: an interpretive argument, or interpretation/use argument (IUA), followed by a validity argument. Kane (2013, 2016, 2021) explained that validation is an ongoing process that (a) starts with outlining the proposed interpretations and uses of test scores in an IUA, which conceptually links test performances to conclusions and decisions based on the test scores, and (b) then evaluates the plausibility of these proposed interpretations in a validity argument. In other words, the IUA outlines the steps validation needs to go through and in what way, whereas the validity argument refers to how well evidence supports or challenges the IUA.

The characteristics of an argument-based approach to validation can be summarized as follows: (a) test developers or researchers specify the various score meanings and uses, (b) claims or inferences made based on the test scores are used as the building blocks to build an IUA, and (c) test developers or researchers use the IUA as a frame for gathering evidence to build the validity argument. This approach provides a systematic process for how validation researchers structure validity arguments and allows researchers or test developers the flexibility to determine the claims they want to make based on test scores and the types of evidence needed to support the claims made, depending on the testing contexts and test uses (Kane, 2013).

The argument-based framework for validation has been employed by many language test developers and researchers to validate both high-stakes (e.g., Chapelle et al., 2008; Schmidgall, 2017) and low-stakes tests (e.g., Chapelle et al., 2010; Chapelle & Voss, 2014; Knoch & Chapelle, 2018; Nguyen et al., 2023; Youn, 2015). Among these, Chapelle et al. (2008) is used as a model for the current paper given its focus on the TOEFL iBT test that measures academic ELP and its comprehensive coverage of evidence collected from early test development to subsequent validation after test launch—a scope of work that is unprecedent in research studies following an argument-based approach to validation. Chapelle et al. (2008) presented a step-by-step illustration of the argument-based framework by first articulating an interpretive argument, which contains a sequence of six inferential steps: (a) domain description, (b) evaluation, (c) generalization, (d) explanation, (e) extrapolation, and (f) utilization. The researchers then specified a list of warrants, assumptions, and evidence to support the score inferences across the six steps. The book chapters present extensive empirical research studies and methods used to collect validity evidence for the assessment of the six inferences in the interpretive argument (see Chapelle et al., 2008, for details). This seminal work provides a holistic and systematic process that guided the development of the TOEFL Junior IUA presented in the next section.

TOEFL Junior Interpretive Argument

To support the adequacy and appropriateness of the TOEFL Junior test scores for intended uses and score interpretations, during test design and development ETS test developers and researchers created a preliminary TOEFL Junior interpretive argument to guide test validation. This high-level framework consisted of the same six inferences proposed in Chapelle et al. (2008) and included brief statements that connected the test scores to their meanings and uses (see So et al., 2015, pp. 22–23). The preliminary framework served as a guide to collecting validity evidence and more than a decade of research has formed an accumulated body of evidence for the validity of the proposed score interpretations and uses.

This paper expands the TOEFL Junior preliminary validation framework by building a TOEFL Junior validity argument, following an argument-based approach to validation (Kane, 2013). Within the context of TOEFL Junior, an interpretive argument was first built where claims about the intended interpretations and uses of the test scores were clearly stated by laying out a network of inferences and their associated assumptions inherent in the proposed interpretations and uses. Second, using the TOEFL Junior interpretive argument as a frame for organizing evidence, the extent to which each claim is upheld with evidence was evaluated in a TOEFL Junior validity argument. The purpose of the validity argument, as explained in Kane (2013, 2016), is to make explicit the meaning of test scores and the basis for claiming various aspects of score meanings and uses.

The proposed TOEFL Junior interpretive argument consists of the following six-step inferences, with assumptions identified for each of the warrants (see Figure 1). Each of these inferences is intended to capture an aspect of score meaning: (a) *domain description*: the test items and tasks represent skills and abilities required for students to have a successful experience in an English-medium secondary school context; (b) *evaluation*: scores on the tests reflect the targeted language abilities and skills; (c) *generalization*: similar test scores are expected to be obtained across different test forms and

Inference: Warrant	*Assumptions (numbered under each warrant)*
F. Utilization: TOEFL Junior tests are useful for assessing students' English language proficiency, monitoring progress, making placement decisions, and supporting teaching and learning.	
↑	1. The meaning of the TOEFL Junior test scores is clearly interpretable by stakeholders. 2. TOEFL Junior test scores are appropriate for intended uses.
E. Extrapolation: Performance on the TOEFL Junior tests reflects students' performance in an English-medium secondary school context.	
↑	1. Performance on the TOEFL Junior tests is related to teacher judgments of students' language abilities. 2. Performance on the TOEFL Junior tests is related to students' self-assessment of English language abilities. 3. Performance on the TOEFL Junior tests is related to other tests of English language proficiency in the target language use (TLU) domain.
D. Explanation: TOEFL Junior scores can be attributed to the test construct of academic English language proficiency in secondary-level education.	
↑	1. Internal structure of TOEFL Junior test scores is consistent with a theoretical view of English language proficiency. 2. The linguistic knowledge, strategies, and processes required to successfully complete TOEFL Junior tasks vary in keeping with theoretical expectations. 3. Test performance varies according to experience in learning English.
C. Generalization: Similar scores on the TOEFL Junior are expected to be obtained across comparable tasks, test forms, and testing conditions.	
↑	1. Sufficient numbers of tasks are included to provide stable estimates of test takers' performances. 2. Scaling and equating procedures are used to ensure score consistency across test forms. 3. The TOEFL Junior test scores are consistent across test forms. 4. Clearly defined test specifications are used to create parallel tasks and forms.
B. Evaluation: Scores on TOEFL Junior reflect the targeted language abilities and skills.	
↑	1. Scoring procedures and guidelines are well-specified and followed. 2. Scoring rubrics are appropriate for judging evidence of the targeted language abilities. 3. Statistical characteristics of test items and tasks are appropriate for providing evidence of the targeted language abilities. 4. Task administration conditions are appropriate for providing evidence of the targeted language abilities.
A. Domain Description: TOEFL Junior test items and tasks represent skills and abilities required for students to have a successful experience in an English-medium secondary school context.	
	1. Assessment tasks produce evidence of abilities representative of language knowledge and skills required for study in an English-medium secondary school context. 2. Teachers and students perceive the test tasks as relevant and representative of the TLU domain.

Figure 1 TOEFL Junior Interpretive Argument

testing occasions; (d) *explanation*: test scores are attributed to a construct of academic English proficiency in secondary level education; (e) *extrapolation*: the construct of academic English proficiency as assessed by TOEFL Junior accounts for students' performance in an English-medium secondary school context; and (f) *utilization*: test scores are useful for assessing students' ELP, tracking students' progress, making placement decisions, and informing teaching and learning. The inferences, warrants, and assumptions start from identifying the TLU domains and tasks as part of the construct definition, move to examining whether the test scores account for relevant abilities from theoretical perspectives, and end with the utilization inference for evaluating decisions based on test scores.

TOEFL Junior Validity Argument

The TOEFL Junior validity argument provides an overall evaluation of the TOEFL Junior interpretive argument to determine how well the available evidence supports the inferences and assumptions in the interpretive argument (i.e., backing evidence) or challenges these inferences (i.e., rebuttal evidence; Kane, 2013). In this paper, validity evidence was collected from three main sources: (a) internal documents regarding operational procedures for item development, statistical analysis of test items, and rater training and calibration; (b) published research studies, and (c) publicly available resources provided by the TOEFL Junior testing program. This body of evidence has been aligned to the assumptions and warrants underlying each inference of the interpretive argument, as discussed in the sections below.

Domain Description

The domain description inference makes the important claim that the TOEFL Junior test items and tasks represent the skills and abilities required for students to have a successful experience in an English-medium secondary school context. This claim is justified if students' performances on the tests reveal relevant knowledge, skills, and abilities in situations representative of those used in the TLU domain. This warrant is further based on the assumption that assessment tasks representing the TLU domain can be identified.

Assumption 1

Assessment tasks produce evidence of abilities representative of language knowledge and skills required for study in an English-medium secondary school context.

Backing

Review of relevant documents and empirical research helped identify and confirm that assessment tasks captured important language tasks that required knowledge and skills used in the TLU domain.

The first backing came from domain analysis and systematic review of various ESL/EFL standards and textbooks during the early stage of test development (So et al., 2015). The test development team reviewed English language standards, curricula, and textbooks used in Brazil, China, France, Korea, Japan, Taiwan, Turkey, and Vietnam. In addition, the team analyzed ELP standards for English learners in U.S. middle schools (i.e., California, Colorado, Florida, New York, and Texas state standards and the WIDA consortium standards) and consulted academic literature on language used in academic contexts. The research team also sought input from experienced classroom teachers and experts in ESL and EFL instruction and gathered insights into language demands required in secondary school contexts. Results of these research efforts helped identify the important real-world tasks and the skills needed to complete those tasks within secondary level education (see detail in So et al., 2015).

Subsequent to the test launch, empirical studies were carried out to examine the appropriateness and relevance of the TOEFL Junior assessment tasks. One such study was a systematic analysis of text complexity in the reading passages used in the TOEFL® Family of Assessments, including TOEFL iBT, TOEFL Junior, and TOEFL Primary® (Chen & Sheehan, 2015). The researchers used an automated text analysis tool, TextEvaluator® (Sheehan et al., 2010), to analyze a collection of reading passages used in these three tests. The TOEFL Junior reading passages were found to have complexity scores within the range specified for secondary school students and were representative of the reading tasks and texts that students would encounter in middle schools in the United States (National Governors Association & Council of Chief State School Officers, 2010). Additionally, Timpe-Laughlin (2018) analyzed the alignment between TOEFL Junior Standard test content and four EFL textbooks used for Grades 7 to 10 in Berlin, Germany. Findings showed that the language knowledge and skills assessed in the test aligned well with those covered in the textbooks.

Assumption 2

Teachers and students perceive the test tasks as relevant and representative of the TLU domain.

Backing

Teachers and students thought that the TOEFL Junior test tasks were relevant to and representative of the language knowledge and skills included in English language classes and curricula.

The first support for this assumption came from empirical investigations of perceived task importance and relevance by teachers. Classroom teachers are key stakeholders of the TOEFL Junior tests because they constitute primary users of the test results; moreover, they have direct knowledge of the target test-taking population and the language knowledge and skills taught in English language curricula and materials. Because the TOEFL Junior tests are intended to be used in contexts where English is taught as a foreign/second/additional language, the perceptions of teachers from these contexts are critical to ensure that the test tasks are relevant and appropriate for measuring the target construct.

During test development, teachers' views about the proposed TOEFL Junior assessment tasks were reported in So (2014). The researcher conducted focus-group interviews with 10 EFL teachers (four from public schools, two from a private school, and four from after-school programs) in South Korea to investigate their perceived relevance and importance of the TOEFL Junior Comprehensive pilot speaking and writing tasks. The teachers thought that the Read Aloud, Picture Narration, and integrated Listen-Speak tasks were good measures of students' speaking skills and relevant to the kinds of activities they used in the classroom. They were generally positive about the Editing, E-mail, Opinion, and integrated Listen-Write tasks but expressed concerns about a piloted Dictation task, suggesting that this task was not a valid measure of students' writing ability. Following additional analyses, the Dictation task was removed, which helped to ensure that TOEFL Junior test tasks were relevant and representative of those in the TLU domain.

In Timpe-Laughlin (2018), the researcher interviewed four teachers from two schools in Berlin, Germany. The teachers indicated that the language knowledge and skills assessed in TOEFL Junior Standard aligned well with the EFL curriculum used. In another study, Galikyan et al. (2019) had 202 students, ages between 11 and 16, from an after-school language program in Armenia take the TOEFL Junior Standard test and respond to a questionnaire that investigated their perceptions of the test design and skills assessed. Nine classroom teachers from the program also reviewed the test materials and responded to a teacher questionnaire that gathered their viewpoints about the test construct and intended uses. Both stakeholder groups reported that the TOEFL Junior Standard test allowed students to accurately demonstrate their English language abilities and that the test tasks reflected the language used in the classroom.

Evaluation

The evaluation inference is based on the warrant that TOEFL Junior test scores reflect the targeted language abilities and skills. This warrant is based on assumptions about scoring, item analyses, and conditions of task administration. Backing in support of these assumptions includes evaluating scoring criteria and procedures, rater training and calibration, running statistical analyses of test items, and comparing performances on different test administration conditions.

Assumption 1

Scoring procedures and guidelines are well specified and followed.

Backing 1

Review of internal documents indicated that standard procedures and guidelines were followed to score test items on the TOEFL Junior Standard test.

Standard procedures for scoring the TOEFL Junior Standard test and for producing the section scale scores are described in internal documents. These documents outline the processes for item analysis, equating, scoring, and scaling (i.e., converting raw score points to scale scores). These procedures are strictly followed by ETS's assessment developers, data analysts, and psychometricians who are involved in scoring the test.

Backing 2

Review of internal documents showed that raters were trained, certified, and calibrated following standard procedures and guidelines.

For the TOEFL Junior Speaking test, internal documents on scoring procedures specify that raters need to complete a training program and pass a certification test using the ETS Online Network for Evaluation (ONE) system. Before each operational scoring session, raters have to pass a calibration test that evaluates their readiness to score for that particular scoring day. As raters score, they are monitored and supported by ETS's scoring leadership team, which provides guidance and feedback to raters when needed to ensure the accuracy and quality of scoring. In addition, rater support documents with benchmark samples of each score point are made available to the raters in the ONE system. Review of these internal documents revealed that the TOEFL Junior raters are thoroughly trained, certified, and monitored in use of the scoring rubrics.

Assumption 2

Scoring rubrics are appropriate for judging evidence of the targeted language abilities.

Backing 1

Rubrics were developed, trialed, and revised based on expert consensus.

Rubric development and pilot testing studies are documented in the TOEFL Junior test design framework paper (So et al., 2015). The scoring rubrics were developed on the basis of best practices for creating appropriate and meaningful scores, results from the pilot studies, and intended uses of the reported scores. The scoring rubrics for the TOEFL Junior speaking and writing measures were developed through an iterative process of rubric development, trialing, and revision.

Backing 2

Research showed that raters are able to score reliably and identify differences in performances across score levels.

So (2014) provides initial empirical evidence concerning issues related to rater reliability. Trained raters double scored TOEFL Junior Comprehensive spoken and written responses, collected from 2,931 test takers who participated in the pilot test. Interrater correlations between the scores given by two raters ranged between .67 and .76 for the speaking items and between .73 and .92 for the writing items, indicating substantial rater agreement.

Ensuing empirical studies using discourse-analytic approaches to examining spoken and written responses have shown that raters were able to apply the scoring rubrics consistently to responses elicited by the test tasks. For speaking, Hsieh and Wang (2019) analyzed features of fluency, grammar, vocabulary, and content quality demonstrated in test takers' responses to the TOEFL Junior Comprehensive Picture Narration and Listen-Speak tasks. Findings showed that the majority of the 21 spoken features examined differentiated test takers across proficiency levels with medium to large effect sizes. For writing, Wolf et al. (2018) analyzed TOEFL Junior Comprehensive test takers' performances on the Opinion writing task and conducted detailed analysis of features of fluency, grammar, vocabulary, and discourse complexity using both human annotation and automated text analysis tools (i.e., TextEvaluator [see Sheehan et al., 2010]) and e-rater (an automated essay scoring system; see Attali & Burstein, 2006, for details of the features included in the system). Results revealed clear relationships between various linguistic accuracy and complexity variables and the students' writing task scores assigned by certified raters. Findings of these two studies provide empirical validity evidence to support the claim that raters are able to differentiate test-taker performances using the scoring rubrics.

It should be noted that, as described earlier, the current TOEFL Junior Writing test is scored using an automated scoring engine that is trained on human scores and covers the same scoring criteria used in the scoring rubrics. The agreement between machine scores and human scores across task types and score levels was evaluated prior to implementation of automated scoring. Pearson correlations between machine and human scores ranged from .77 to .80 across tasks, similar to correlations between two human raters, which ranged from .73 to .80. Human-machine score agreement was also evaluated across test takers from different language groups to ensure consistent scoring and to ensure agreement in classification of test-taker proficiency in terms of the CEFR levels. The results provided evidence supporting the reliability of using machine scoring for the TOEFL Junior Writing test.

Assumption 3

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

Backing

Item analyses were conducted to verify that the difficulty and discrimination of test items and tasks were appropriate.

The psychometric quality of the TOEFL Junior test items and tasks was examined throughout test development. For example, So (2014) reported analyzing item difficulty on the TOEFL Junior Comprehensive pilot speaking and writing tasks to determine if the test tasks produced appropriate levels of difficulty and differentiated among test takers' levels of ability. Young et al. (2013) conducted a differential item functioning (DIF) analysis on the TOEFL Junior Standard pilot test forms and found construct-relevant explanations for the items that exhibited significant group differences in the DIF analysis. Overall, results of item analyses based on the TOEFL Junior pilot data provide good backing for the psychometric quality of the test scores during early stages of test development.

After the test launch, the TOEFL Junior data analysts and psychometricians conduct routine analysis of item characteristics, following ETS Standards for Quality and Fairness (ETS, 2014). The TOEFL Junior 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, 2022c).

Assumption 4

Task administration conditions are appropriate for providing evidence of the targeted language abilities.

Backing

Empirical research findings showed that task administration conditions were appropriate.

Backing for the assumption that the task administration conditions were appropriate was initially derived from prototyping and pilot studies around the world. These studies focused on whether the young test takers were able to understand the test instructions and complete the test items or tasks in a manner that was appropriate given their language abilities and age. So (2014) reported that, through the initial trials of the TOEFL Junior Comprehensive speaking and writing test tasks, one speaking and one writing task were eliminated because of concerns about testing time for young learners, coupled with construct relevance and psychometric quality of the tasks. The amount of time required for completing the test tasks was finalized based on results of the prototyping and pilot testing studies.

External researchers have further explored other conditions for test administration and provided evidence that the TOEFL Junior test administration conditions are appropriate for measuring the English language ability of young learners with different backgrounds or needs. For example, for the Listening Comprehension section of the TOEFL Junior Standard test, Eberharter et al. (2023) compared the standard single-listening administration mode with a self-paced mode. The participants were 139 eighth-grade EFL students from four secondary schools in Austria who had reading-related learning difficulties. Each student completed 15 items in a single-listening condition and another 15 in a self-paced condition. Analysis of the students' performances yielded comparable test scores between the two conditions, supporting valid score meaning and interpretation of the standard single-listening administration condition for this group of learners. In another EFL context, Yeom and Jun (2020) had 84 seventh to ninth graders from four public schools in South Korea take the paper- and computer-delivered TOEFL Junior Standard Reading Comprehension test items. Analysis of the students' performances resulted in equivalent scores between the two test-delivery modes. Findings of the two studies provided backing to support the appropriateness of the TOEFL Junior Standard test administration conditions for measuring the targeted language abilities.

Generalization

The generalization inference is based on the warrant that similar test scores are expected to be obtained across comparable tasks, test forms, and testing conditions. Two assumptions underlie this warrant and concern issues of reliability and meaningful and consistent score interpretation. Backing for the assumptions was obtained through studies of generalizability and reliability and through rigorous task design and development.

Assumption 1

Sufficient numbers of tasks are included to provide stable estimates of test takers' performances.

Backing

Results from reliability and generalizability studies indicated the number of test tasks was appropriate to maintain required levels of reliability.

So (2014) provided evidence concerning issues related to the number of speaking and writing tasks required to maintain reliability for the TOEFL Junior Comprehensive speaking and writing test sections. Generalizability theory (G-theory) analyses on pilot test items were used to examine the influence of reducing one speaking and one writing test item on the reliability of the test while also considering teachers' feedback about the relevance and importance of the proposed pilot test tasks as described earlier. Results of the G-theory analyses showed that removing one speaking and one writing item did not have a substantial impact on the reliability of the TOEFL Junior Comprehensive speaking and writing test sections. For speaking, the G-coefficient was .92 for five double-scored items and dropped to .90 for four double-scored items. For writing, the G-coefficient was .88 for five double-scored items and dropped to .85 for four double-scored items. The research team considered that the generalizability coefficients were acceptable for the intended uses of the test and included four speaking and four writing items in the final test specifications.

Assumption 2

Scaling and equating procedures are used to ensure score consistency across test forms.

Backing

Review of relevant documents and publications provided evidence that procedures for scaling and equating are well specified and adhered to.

This assumption is supported by an internal document that outlines the process of ongoing scaling and equating of test sections and forms. Scaling and equating of the test forms is carried out using an item response theory framework to estimate item parameters. For each of the TOEFL Junior Standard test section, the total number of correct selected-response items (i.e., 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 used interchangeably. The equated raw scores are then converted to section scale scores that range from 200 to 300. Scores on any new TOEFL Junior Standard test form are equated and then reported on a common scale. Scaling and equating procedures ensure consistent interpretation of scores across test forms and administrations.

Assumption 3

The TOEFL Junior test scores are consistent across test forms.

Backing 1

The reliability coefficients and standard error of measurement (SEM) of the TOEFL Junior tests indicate that the test scores are consistent across test forms.

The two statistics commonly used to describe the reliability of test scores of a group of test takers are the reliability coefficient and the SEM. The reliability coefficient is an estimate of the correlation between scores on different test forms and indicates the degree to which a test form produces consistent scores. It varies from .00, indicating no agreement at all, to 1.00, indicating perfect agreement. The reliability coefficients for the TOEFL Junior Standard scale scores are .87 for Listening Comprehension, .87 for Language Form and Meaning, .89 for Reading Comprehension, and .95 for the total test, showing adequately high score consistency across test forms (ETS, 2018, p. 24).

The SEM indicates the extent to which test takers' scores differ from their true scores. A test taker's "true score" is the average of the scores that a given test taker would earn on all possible test forms. The difference between a test taker's true score and the score the test taker actually earns is called the "error of measurement." The SEM, for a group of test takers, is the average size of those differences. The SEM for each of the TOEFL Junior Standard test section scale scores is 9.8 for Listening Comprehension, 9.0 for Language Form and Meaning, and 10.0 for Reading Comprehension, and the SEM for the total test score is 16.6 (ETS, 2018, p. 24).

For the TOEFL Junior Speaking and Writing tests, raw scores are reported to test takers because the meanings of the raw scores can be more easily interpreted through the descriptions in the scoring rubrics and thus the scores are not equated (So et al., 2015). Score comparability of the TOEFL Junior Speaking and Writing tests is maintained through trying out new test items in small-scale sessions before they are used operationally as well as through rigorous rater training and monitoring. The reliability coefficient for the TOEFL Junior Speaking test is .87 and the SEM is 1.24 (ETS, 2018, p. 31). The reliability coefficient and SEM for the newly launched TOEFL Junior Writing test is not available at the time of writing and will be reported when available.

Backing 2

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

Empirical evidence pertaining to the stability or test-retest reliability for the TOEFL Junior Standard test came from Gu et al. (2015), who conducted an analysis of repeat test takers' performances. In a subset of the data, performances of 619 test takers who had taken the test twice within a 2.5-month period showed that all three mean test section scale score changes (e.g., 2.5 score point on Listening Comprehension) were smaller than the SEM for each test section. These operational test data provide evidence supporting the stability of the test scale scores across administrations.

Assumption 4

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

Backing

TOEFL Junior item specifications are used for producing parallel items and tasks.

The assumption is backed by the TOEFL Junior item writers' use of well-defined item specifications. Confidential item specifications provide detailed descriptions of the test tasks and identify which aspects of the test tasks must remain constant to consistently assess pertinent language knowledge and skills and which aspects can be allowed to vary to create new items (e.g., topics for the speaking tasks). Checklists are included in the specifications such that item writers can double check and ensure that the standard steps involved in item development are followed and completed.

Explanation

The explanation inference is based on the warrant that the TOEFL Junior test scores can be attributed to the test construct of academic ELP in secondary-level education. Sources of backing for this warrant include investigations of the internal structure of the tests, the linguistic knowledge, processes, and strategies underlying task performance, and the relationship between test performance and experience in learning English.

Assumption 1

Internal structure of TOEFL Junior test scores is consistent with a theoretical understanding of academic ELP.

Backing

Expected correlations were found among measures within the TOEFL Junior tests and the factor structure of measures was consistent with theorized relationships.

Backing for the assumption included factor analysis studies conducted to support a theorized factor structure of the TOEFL Junior tests. Gu (2015) analyzed the TOEFL Junior Comprehensive pilot test data collected from 436 participants in 15 countries to investigate the latent structure of English language ability in EFL school-age learners. The results showed that the learners' test performances were best explained by a higher-order factor model, representing a global English language ability and four first-order factors that corresponded to the four language skills. Building on the work of Gu, Manna et al. (2018) investigated the factor structure of the English language abilities measured by the TOEFL Junior

Comprehensive test. Data were collected from 2,885 secondary school students in Japan to control for variations in first language and academic context. Analyses of the data showed that a correlated four-factor model associated with the four language skills best represented the English language abilities measured by the test. Though differing in the final selected latent factor models, the two studies provide validity evidence to support the four-skills structure of the ELP measured by the TOEFL Junior tests and support the theoretical view that ELP is composed of highly interrelated components (Bachman & Palmer, 2010).

Assumption 2

The linguistic knowledge, strategies, and processes required to successfully complete TOEFL Junior tasks vary in keeping with theoretical expectations.

Backing 1

Empirical studies showed that the linguistic knowledge and skills required to complete the TOEFL Junior tasks aligned with expected developmental patterns and theoretical definitions of language proficiency.

The first source of backing came from empirical studies that examined the linguistic knowledge and skills required to complete the TOEFL Junior speaking and writing tasks. For the speaking tasks, backing is found in Gu and Hsieh (2019) and Hsieh and Wang (2019), who analyzed features of fluency, vocabulary, grammar, and content quality exhibited in TOEFL Junior Comprehensive test takers' spoken responses. Results of the two studies converged, demonstrating that responses receiving increasing higher scores showed expected developmental patterns in speaking proficiency phenomena specified in theoretical models of communicative competence (e.g., Bachman & Palmer, 1996). These phenomena included construct-relevant dimensions such as increasing speech fluency, grammatical complexity, and content quality (e.g., Frost et al., 2011; Iwashita et al., 2008).

For the writing tasks, backing was found in Wolf et al. (2018), who conducted detailed analyses of test takers' responses on the TOEFL Junior Comprehensive Opinion writing task. The responses were analyzed for aspects of essay length, lexical complexity, grammatical accuracy, syntactic complexity, and discourse complexity. The written features analyzed varied as expected with scores assigned by trained raters and were consistent with expected characteristics of argumentative writing (e.g., Cumming et al., 2005; Grant & Ginther, 2000).

Further backing was found in Hsieh (2023), who examined the association between L2 reading and writing skills using measures of the TOEFL Junior tests. The participants included 185 students in Grades 7 and 8 from Denmark, Finland, and the Netherlands who completed the TOEFL Junior Standard test. The researcher also collected the students' written responses to the TOEFL Junior Email writing task and a researcher-developed descriptive writing task. Findings showed that the students' TOEFL Junior Standard reading comprehension test scores were related to their writing performances, particularly in terms of writing fluency (amount of text produced) and idea development. More advanced readers showed a greater level of linguistic sophistication when producing the written texts. This link between reading and writing reflects the theoretical expectation of an association between reading-writing skills, as found in first language (L1) literacy research (Fitzgerald & Shanahan, 2000).

The relationship between TOEFL Junior test performances and theoretical expectations for the language knowledge and skills required for completing the test tasks is further supported by several research studies funded by the TOEFL Committee of Examiners and TOEFL Young Students research grants (e.g., Kim, 2023; Kim et al., 2022; Wallace, 2020, 2021; Wallace & Lee, 2020). For example, a concern that second language (L2) listening performance was jointly determined by both linguistic knowledge (e.g., vocabulary size) and nonlinguistic factors (e.g., executive functions) was tested in Wallace and Lee (2020). The researchers had 209 Japanese senior high school EFL learners respond to a practice version of the TOEFL Junior Standard Listening Comprehension section, a listening vocabulary levels test, and several tests that measured the students' executive function in terms of working memory (e.g., revising information held in temporary storage, switching attentional focus among mental representations). Findings revealed that the students' vocabulary knowledge was the only variable that significantly predicted their listening comprehension scores. This result supported the claim that language ability—the target construct of the TOEFL Junior tests—should be the main factor influencing successful test task completion, whereas recognizing that many peripheral factors such as cognitive ability, background knowledge, strategic competence may contribute to a test taker's success (Hulstijn, 2015; So et al., 2015).

Similarly, Kim et al. (2022) investigated the extent to which L2 grammar and vocabulary knowledge and cognitive factors (i.e., working memory capacity (WMC), L1 inferencing ability) predicted L2 listening comprehension for passages of different lengths. The participants were 193 ninth grade EFL students in South Korea who responded to a research version of the TOEFL Junior Standard test, and the instruments measured WMC and L1 inferencing ability. Outcomes of the study demonstrated that the students' linguistic knowledge, as measured by performance on the Language Form and Meaning section of the TOEFL Junior Standard test, showed strong links to their listening comprehension for both shorter and longer listening passages. Conversely, WMC was weakly associated with comprehension of longer passages, and L1 inferencing ability was weakly associated with comprehension of shorter passages. Given that WMC and L1 inferencing ability are peripheral to the test construct, the fact that they played minimal roles in the students' listening comprehension performance provides another piece of validity evidence, implying that students' listening performance is related to the expected language knowledge and skills required for comprehension (Buck, 2001).

Collectively, this body of empirical studies provides strong evidence that the TOEFL Junior test scores are influenced by relevant L2 linguistic knowledge and skills required to successfully complete the test tasks and are consistent with the theoretical models of language proficiency on which the TOEFL Junior tests are based (Bachman & Palmer, 2010).

Backing 2

Task completion strategies and cognitive processes used by test takers were consistent with expected processes for successfully completing the TOEFL Junior tasks.

Backing for this assumption was found in studies that examined strategy use and cognitive processes test takers engaged in when responding to the TOEFL Junior test tasks. Hsieh and Gu (2020) collected verbal report data from 31 students in Grades 4 to 6 of an English-medium primary school in Hong Kong to explore the students' strategy use when responding to two TOEFL Primary and one TOEFL Junior picture-based speaking tasks. After the students responded to each task, they reported or reflected on the processes of deriving their spoken responses. Analyses of the verbal reports indicated that the students relied on construct-relevant organizational and elaboration strategies when responding to the TOEFL Junior Picture Narration task.

Similarly, Choi and Loewen (2022) explored task-specific strategic behaviors of 45 EFL learners in upper primary and lower secondary schools in South Korea. The students completed two TOEFL Junior Picture Narration and two integrated Listen-Speak tasks and participated in stimulated recall sessions to report their strategy use when responding to the tasks. Results showed that the students utilized a variety of cognitive and metacognitive strategies that were relevant to the tasks or test construct (e.g., evaluating the content, planning). This study also documented task-specific strategy use, in line with theoretical expectations and empirical work based on adult L2 learners (Swain et al., 2009).

Yeom and Jun (2020) investigated strategy use in reading comprehension. The participants included 84 EFL learners in Grades 7 to 9 in South Korea who completed selected TOEFL Junior Standard Reading Comprehension test questions and filled out a questionnaire of reading and test-taking strategies. Findings suggested that students with different levels of English reading proficiency differed in their strategy use. High-proficiency readers reported using more strategies to enhance the quality of their reading comprehension (e.g., use logical connectors to clarify content and passage organization) than the lower-proficiency readers, as expected.

With respect to the cognitive processes underlying test performance, most empirical studies investigated the role of WMC, along with linguistic knowledge required to complete test tasks as discussed above. The impact of WMC on test takers' reading comprehension and writing performances were investigated in two studies that employed the same data set collected from 94 students in Grades 6 and 7 who were from two English-medium schools in Hungary. The first study, Michel et al. (2019), reported the impact of WMC on the students' performance on the TOEFL Junior Comprehensive Email and integrated Listen-Write tasks. The second study, Brunfaut et al. (2021), focused on the effects of WMC on students' performances on the TOEFL Junior Comprehensive Reading Comprehension section. Findings of the two studies converged, suggesting that WMC had limited effects on reading comprehension and writing performance. Given that WMC is peripheral to the language ability construct measured by the TOEFL Junior tests, these studies provide evidence that young learners' varying levels of working memory functions do not cause construct-irrelevant variance in TOEFL Junior test scores.

Affective factors, such as task motivation or test anxiety, that test takers experience during task completion can also influence test performance, especially for young learners whose ability to exercise control over their emotional status

is still undergoing development (Butler, 2017). Kormos et al. (2020) investigated the impact of task motivation on test performance among 104 Hungarian primary school students, ages 11 to 15. The students completed the TOEFL Junior Comprehensive integrated Listen-Speak and Listen-Write tasks and then filled out a task-motivation questionnaire that assessed their posttest feelings, reported effort, task appraisal, emotional state, and result assessment. Findings revealed that the task-motivational variables (e.g., task effort, task anxiety) were largely unrelated to the students' task performances. The researchers attributed the findings to positive reactions from the students, resulting from appropriate task administration conditions, as well as the authentic task design that reflected the English-medium instructional context that the participating students were engaged in. The study results also indicated that variations in the students' motivation did not create construct-irrelevant variance, providing validity evidence to support score interpretation.

Assumption 3

Test performance varies according to experience in learning English.

Backing

Research results showed that test performance improved with greater experience in learning English.

Backing for this assumption came from research comparing differences in TOEFL Junior test scores among learners with varying degrees of experience in learning English. Based on students' repeated performances on the TOEFL Junior Standard test, Madyarov et al. (2021) found that the test scores of beginner-level Armenian adolescent students, ages 11 to 17, in an after-school English language program improved significantly after receiving 20 hours of instruction over a 10-week period. Ling and Gu (2019) analyzed operational test data and self-reported background information to examine the associations between TOEFL Junior Comprehensive test scores and construct-relevant learning experiences. Results of the analysis showed that a greater number of years studying English and a longer length of stay in an English-speaking country were significantly associated with higher levels of ELP as measured by the TOEFL Junior Comprehensive.

Also using self-reported background information, Huang et al. (2021) used a survey to gather information about participating students' hours of formal English instruction received, onset age of learning, and frequency of out-of-school contact with English. The correlations between the students' test scores on the TOEFL Junior speaking practice items and the construct-relevant learning experience variables were moderately high (between .30 and .49). Findings of the three studies collectively provide evidence to support the claim that TOEFL Junior test scores differ based on the amount and quality of experience in learning English. These findings are also consistent with the differences that one would expect based on a theoretical construct of academic language proficiency that develops with time spent learning English (Chapelle et al., 2008).

Extrapolation

The extrapolation inference is based on the warrant that performance on the TOEFL Junior tests reflects students' performance in an English-medium secondary school context. Underlying this warrant is the assumption that test-taker performance is related to other criteria of language proficiency in the TLU domain. Evidence relevant to this warrant comes from research studies investigating variables in the TLU domain such as teacher judgements of students' language abilities and test takers' self-assessments or about the relationships between TOEFL Junior test scores and other external measures of English.

Assumption 1

Performance on the TOEFL Junior tests is related to teacher judgments of students' language abilities.

Backing

Empirical studies showed that TOEFL Junior test scores were related to teacher judgments of students' language abilities.

Papageorgiou and Cho (2014) examined the relationship between students' test scores on the TOEFL Junior Standard test and teachers' evaluation of students' English language ability related to course placement decisions. The TOEFL Junior

Standard test was administered to 92 students in Grades 7 to 9 from two secondary schools that offered varying levels of ESL courses. The researchers also asked classroom teachers to evaluate the accuracy of each student's ESL placement and recommend an appropriate ESL level if they thought that the students' current placement was inaccurate based on their knowledge of the students' language abilities. Strong correlations between the students' test scores and teachers' judgments were found in both schools ($r = .72$ and $r = .83$, respectively).

Huang et al. (2021) also used teacher judgments as an indicator of students' ELP in a secondary school context in Taiwan. Six English teachers were recruited to evaluate their students' English language abilities using a 100-point scale. In addition, 252 students in Grades 7 to 9 from these teachers' classes completed the practice version of the TOEFL Junior Comprehensive speaking test. The correlation between the students' speaking test scores and teacher ratings of students' speaking proficiency was $r = .62$. Results of these two studies provide empirical evidence in support of the links between students' performances on the TOEFL Junior tests and teachers' judgments of students' English language abilities.

Assumption 2

Performance on the TOEFL Junior tests is related to students' self-assessment of English language abilities.

Backing

Empirical studies found that TOEFL Junior test scores were associated with students' self-assessment of English language abilities.

In addition to teacher ratings described above, Huang et al. (2021) used self-assessment as another nontest criterion of students' English language ability. The participating students were asked to evaluate their own English listening, speaking, reading, and writing skills using a 100-point percentile scale based on how they compared to other students of the same age. A moderate correlation was observed between students' self-assessment and their scores on the speaking practice test ($r = .57$). In another study that employed self-assessment, Kormos et al. (2020) had students evaluate their performances on the TOEFL Junior Comprehensive integrated Listen-Speak and Listen-Write tasks using a 4-point scale. The students' subjective evaluations were significantly, though weakly, correlated with their performances on the Listen-Speak ($r = .20$) and Listen-Write ($r = .21$) tasks.

As Chapelle et al. (2008) pointed out, high correlations of self-assessments with test scores are not expected due to the different measurement methods used and the differences in the constructs of perceived language ability and test performance. Thus, for both studies, the moderate to weak correlations observed provide validity evidence for the assumption that TOEFL Junior test scores are, as expected, related to self-assessment of language proficiency in secondary school contexts.

Assumption 3

Performance on the TOEFL Junior tests is related to other tests of ELP in the TLU domain.

Backing

Research results showed positive relationships between TOEFL Junior test scores and other tests of ELP in secondary school contexts.

The validity evidence for the relationship to other criteria of ELP in the TLU domain refers to the degree to which TOEFL Junior test scores are similar to the scores of other tests that measure similar constructs (Messick, 1996). Backing of this assumption was obtained through correlations that were estimated between the scores of the TOEFL Junior tests and other measures of academic English proficiency appropriate for the target population.

Kamiya (2017) collected data from 144 students in Grade 12 who took the English section of the National Center Test (NCT), a national matriculation exam used in Japan for college admissions, and the TOEFL Junior Comprehensive test. The NCT English test measured similar constructs of academic English reading and listening skills. The correlations between the students' scores on the NCT listening and reading tests and the TOEFL Junior Comprehensive reading and listening sections were between .71 and .78.

As mentioned earlier in the paper, Huang et al. (2021) had 252 students in Grades 7 to 9 in Taiwan respond to four TOEFL Junior speaking practice tasks. The students also completed two construct-related speaking monologue tasks and two interactive speaking tasks that were developed by the researchers and validated in a larger assessment development project (Bailey & Heritage, 2014). The correlations between the scores on the TOEFL Junior speaking practice tasks and the researcher-developed speaking tasks ranged from .55 to .73. Results of the two empirical studies demonstrate that TOEFL Junior test scores are related to construct-relevant external measures of ELP.

Utilization

The utilization inference is based on the warrant that TOEFL Junior test scores are useful for measuring students' ELP, tracking progress, making placement decisions, and supporting teaching and learning. This warrant relies on the assumptions that the test scores are clearly interpretable by stakeholders and are appropriate for intended uses.

Assumption 1

The meaning of TOEFL Junior test scores is clearly interpretable by stakeholders.

Backing 1

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

Test developers and researchers at ETS have conducted a series of studies to develop score levels and descriptors that provide meaningful information to test takers, parents, teachers, and score users. For example, to enhance score meanings and interpretability of the TOEFL Junior Standard total scale scores, Papageorgiou, Morgan, and Becker (2015) developed performance levels and descriptors for score reporting purposes. Data were collected from 3,607 students who took an operational test form in 2012. The researchers analyzed the accuracy and consistency of classifying the test takers into four, five, and six levels, defined by different cut scores. Considering the intended uses of the test, ultimately a five-level solution was selected for reporting overall test performance. Performance level descriptors were then created by analyzing test-taker performance, item difficulty, and the relevant CEFR level descriptors (Council of Europe, 2001). In the TOEFL Junior Standard test score report, the typical profile of each performance level on the test is also reported in terms of CEFR levels to support score interpretations for users familiar with the CEFR.

Similar research and development effort for facilitating score interpretability was made for the TOEFL Junior Comprehensive test. Papageorgiou, Xi, et al. (2015) developed band levels for overall score and accompanying performance descriptors. Data were collected from 2,931 pilot test takers, which were used in identifying six band levels and creating performance level descriptors for reporting purposes. Multiple sources of information were drawn upon in this effort, including test-taker performance on the test, speaking and writing scoring rubrics, characteristics of the test items, typical student performance profiles, the performance of norm groups on the test, and relevant CEFR level descriptors.

To help stakeholders understand score meanings and make score-based decisions, the TOEFL Junior testing program also offers information about the test constructs targeted by the TOEFL Junior tests and the meanings of test scores, as well as sample score reports, sample test items, and performance descriptors. These materials are located on the TOEFL Junior Website (https://www.ets.org/toefl/junior.html) and in test-taker handbooks (ETS, 2018, 2023). 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.

Backing 2

TOEFL Junior test scores are linked to external language proficiency standards or frameworks to enhance score meanings and facilitate interpretation.

ETS researchers and external research collaborators have conducted several standard setting and score mapping studies to link the TOEFL Junior test scores to external language proficiency standards and frameworks. These mapping results guide stakeholders in understanding students' abilities relative to widely accepted international standards with the goal of supporting teaching and learning worldwide. For example, Baron and Tannenbaum (2011) and Tannenbaum and Baron (2015) carried out standard setting studies linking the TOEFL Junior Standard and TOEFL Junior Comprehensive

tests, respectively, to CEFR levels. Additionally, Baron and Papageorgiou (2016) conducted a standard setting study where recommended cut scores for ESL placement decisions were identified for the TOEFL Junior tests. In these studies, language education experts and teachers were carefully recruited from different types of schools or ESL programs around the world to ensure proper representation. Rigorous procedures and established standard setting methods, including the modified Angoff procedure (Cizek & Bunch, 2007) and Performance Profile method (Fleckenstein et al., 2020; Hambleton et al., 2000; Zieky et al., 2008), were systematically followed during the standard setting meetings. The researchers also collected evidence addressing the validity of the standard setting processes by gathering panelists' perceptions on their own standard setting judgments.

Additionally, as described earlier in the paper, the TOEFL Junior Standard Reading Comprehension section scale scores have been mapped to the Lexile Framework for Reading, which places the ability of the reader and the difficulty of the texts on the same scale. The TOEFL Junior Standard test score report includes the Lexile measure to help students choose books at the right reading level to improve their English reading proficiency.

More recently, in a joint project between ETS and the National Education Examinations Authority (NEEA; Ministry of Education, China), researchers at ETS and NEEA conducted a study to map the scores of the TOEFL Junior Standard and TOEFL Junior Speaking tests onto the China's Standards of English (CSE; NEEA, 2018). To estalish content alignment, or construct congruence, between the tests and CSE levels, the research team first analyzed the CSE level descriptors to identify those that were relevant to the TOEFL Junior test content. A standard setting meeting was then held with 16 language educators in China who represented a variety of Chinese institutions involved in teaching the age groups targeted by the TOEFL Junior tests. T he study was published in a bilingual Chinese and English journal (see Papageorgiou et al., 2022) to help disseminate the score mapping results to Chinese educators and policy makers and to inform language teaching and assessment in the local educational contexts.

Assumption 2

TOEFL Junior test scores are appropriate for intended uses.

Backing 1

Students and teachers perceived the TOEFL Junior tests to be suitable for intended uses.

Galikyan et al. (2019) sought feedback from students and teachers regarding the usefulness of the TOEFL Junior Standard test within the context of an after-school program in Armenia. Results of survey responses indicated that both the students and teachers considered that the TOEFL Junior Standard test was accurate in measuring students' English language skills. The teachers also thought that the test could be an effective tool for capturing changes in students' ELP over time and helping them decide whether a student was ready to move on to the next course level. As described earlier, Timpe-Laughlin (2018) interviewed four EFL teachers in Germany. These teachers had been using the TOEFL Junior Standard test in their classes and were positive about the use of the test to measure their students' English language abilities and monitor students' progress over time as well as to promote students' learning motivation.

Backing 2

Empirical studies showed that TOEFL Junior test scores effectively reflected learning gains and could be used for progress monitoring.

Gu et al. (2015) analyzed 4,600 repeat test takers' performances on the TOEFL Junior Standard test to evaluate the extent to which the test scores were consistent with changes in underlying language abilities resulting from English language learning. Statistical modeling, using the time interval between test administrations as an indicator of language growth given that test takers were actively studying English, revealed that test takers with longer intervals between retesting (e.g., longer than 250 days) showed greater score gains than did those who retested at shorter intervals (e.g., shorter than 75 days). The study provides empirical support for the claim that the TOEFL Junior Standard test can document changes in language ability and be used for monitoring growth over time.

Adding to this empirical evidence, Madyarov et al. (2021) investigated the use of TOEFL Junior Standard test as a measure of progress for students enrolled in an after-school language learning program. The researchers collected 154 students'

performances on the test on three occasions, at intervals of 10 and then 20 instructional weeks within a single program to reduce the variability of learning experiences among the students—an issue not controlled for in Gu et al. (2015). They found that the test was sensitive to learning gains for learners at the A1–A2 CEFR levels who received 20 instructional hours over a 10-week period. However, the test was not as sensitive to learning gains for B1-B2 level students. The finding corroborates previous research that suggests that language learners from higher levels of proficiency may take longer to show score gains measurable by standardized tests (e.g., Elder & O'Loughlin, 2003). This means that the TOEFL Junior Standard test may be a more useful tool for monitoring progress for lower level learners, though future validity evidence is needed to investigate this hypothesis.

Backing 3

The TOEFL Junior tests had positive washback on language teaching and learning.

Wolf et al. (2023) conducted a longitudinal study over 2 years to investigate the washback from using the TOEFL Young Student Series (YSS) tests, including TOEFL Primary and TOEFL Junior, within five schools in Turkey. The researchers used multiple research methodologies, including interviews of school administrators and classroom teachers, analysis of instructional logs from participating teachers, analyses of textbooks used, and surveys of teachers, administrators, parents, and Grades 3 to 7 students. They also gathered students' performances on the TOEFL YSS tests over the 2-year period. Analyses of the multiple sources of data revealed that the use of the TOEFL YSS tests had limited washback effect at the micro-level in daily classroom instruction, with no observable changes to the teaching content and methods resulting from the test use. Interestingly, the researchers observed some positive macro-level washback effects. For example, two schools used the test scores to identify students who needed extra support and offered after-school instruction to those in need. Other macro-level washback included increased student motivation to learn so as to succeed on the TOEFL YSS tests. However, the teachers and parents reported having difficulties understanding the scale scores of the TOEFL YSS tests, especially the TOEFL Primary test scores. While the challenges commented were mostly related to the TOEFL Primary test, the finding indicates that additional support or resources are needed to help stakeholders better interpret and use the test results.

Discussion

The purpose of building a validity argument for the TOEFL Junior tests was to collect validity evidence to support claims regarding the interpretations and uses of test scores for intended purposes. Overall, the validity argument provides systematic support for these claims. Despite the low-stakes nature of the TOEFL Junior tests, extensive documentation and empirical studies exist to support each of the six major inferences outlined in the interpretive argument. This body of evidence supports the conclusion that TOEFL Junior tests are appropriate for measuring students' ELP, tracking students' progress over time, making placement decisions, and supporting teaching and learning, though further backing regarding positive washback and the use of automated scoring for the TOEFL Junior Writing test is required. This section summarizes and assesses the main findings in relation to the inferences in the interpretive argument.

The first inference in the argument, domain description, was fully supported by several domain analysis activities carried out during test development that informed task design and through ensuing empirical investigations of perceived task relevance and importance by teachers and students. These stakeholder perceptions provide important evidence for the claim that assessment tasks are representative of the instructional domain targeted by the test.

For the evaluation inference, the validity evidence shows that the test items and scoring procedures are adequate for intended interpretations. This inference was backed by evidence gathered from internal documents regarding routine procedures and guidelines for operation of the test, including statistical analyses of test items, producing test scores, rater training and certification, and empirical examination of different test administration conditions. Results of these reviews and empirical studies provide substantial evidence to support the claim that the TOEFL Junior test scores accurately reflect intended language skills and abilities.

The generalization inference requires evidence to show that test takers would receive consistent scores on comparable assessment tasks and test forms and across test administrations and scoring conditions. Evidence from a variety of sources provide the needed validity evidence supporting the consistency of test scores and score meanings. Such sources include results of generalizability and reliability studies, information provided in the internal documents regarding scaling and

equating procedures, and the use of item specifications to create parallel forms. However, a reliability estimate for the newly launched TOEFL Junior Writing test is not yet available at the time of writing and needs to be provided when available to show backing for the consistency of scores on this test. Additional research is also needed to evaluate the use of AI capabilities to score the TOEFL Junior Writing responses to ensure that the machine-generated scores are generalizable across tasks, test forms, and testing conditions. This topic is particularly important given that issues regarding the use of automated scoring of young learners' essays, such as difficulty in detecting off-topic or gibberish responses and score consistency across subgroups of learners (e.g., across grade level), have been raised in a recent study that investigated machine-scoring of student essays for Grades 3 to 6 (see Hannah et al., 2023). Another topic worth noting pertains to the application of AI technology to provide more fine-grained feedback to YLLs and teachers. The feasibility and validity of incorporating AI-generated feedback in the TOEFL Junior Writing score report deserve research and development attention to further support positive impacts on language teaching and learning — one of the main uses of the TOEFL Junior tests.

The explanation inference links the test items and scores back to the test construct definition. Kane (2013) argued that a test that involves making inferences about a theoretical construct would require stronger evidence to support the link between test scores and the theory than those that do not make such claims. As stated in So et al. (2015), Bachman and Palmer's (2010) model of language knowledge serves as the theoretical framework for the design of the TOEFL Junior tests and informs the link between the assessment tasks and the TLU tasks. To evaluate the link to theoretical expectations of task performances on the TOEFL Junior tests, extensive empirical studies were carried out that examined test response characteristics, the linguistic knowledge, processes, and strategies required to perform the test tasks, the internal structure of the tests, and the relationships between the tests and external measures of similar constructs. Findings of the collected empirical studies provide substantial backing for the claim that TOEFL Junior test scores are attributable to the theoretical construct of academic ELP in secondary education.

The extrapolation inference links the test scores to claims about a test taker's language knowledge and ability in the TLU domain, outside of the testing setting. Evidence was gathered to show that test-taker performance is related to other criteria of language proficiency in the TLU domain. This validity evidence includes the relationships between TOEFL Junior test scores and scores on researcher-developed speaking and writing tasks, teacher judgements of students' abilities, and students' self-assessment. Although the extrapolation inference is adequately supported for the intended score interpretations, relatively fewer empirical studies exist to support the assumptions underlying this inference. Future studies might explore additional measures of English proficiency (e.g., classroom-based assessments) that would enable meaningful comparisons between students' test performance and their language performance in the TLU domain.

Finally, the utilization inference relates to score interpretation and, more importantly, uses, which in part involves consideration of how test scores and uses are communicated to stakeholders. The meanings of the TOEFL Junior test scores are communicated to test takers, teachers, parents, and score users in score reports, test taker handbooks, the TOEFL Junior website, and research reports or publications. These materials help promote effective and appropriate uses of the tests and the avoidance of misuse. In addition, the usefulness of the TOEFL Junior test scores were investigated in research studies that evaluated perceptions of test use in different language learning contexts, and the feasibility of using the tests for monitoring progress. These studies provide substantial support for the claim that the tests are helpful for uses such as measuring students' ELP, tracking progress, and informing language teaching and learning. Nonetheless, despite the abundance of evidence supporting the utilization inference, as revealed in Wolf et al. (2023), limited direct evidence is available regarding the impact of the TOEFL Junior tests on English teaching. Future research endeavors are warranted to explore factors that influence washback in the language classrooms and how classroom teachers can better utilize the test results to effectively promote teaching and learning. In addition, some stakeholders reported having difficulty understanding the scale scores in Wolf et al. In light of the research findings, the TOEFL Junior testing program may consider offering additional score interpretation materials to enhance the comprehensibility of score reporting information. Empirical investigations regarding how stakeholders interpret and *actually* use the score reports will provide an additional piece of important validity evidence to support the utilization inference.

Conclusion

This report is one of only a few examples of the use of an argument-based framework to validating language assessments designed for YLLs. This document adds to the existing body of research supporting the TOEFL Junior tests by articulating the full range of an interpretive argument for the tests and collecting comprehensive validity evidence to evaluate

the proposed interpretive argument. To the best of my knowledge, this report is the first of its kind for a young learner assessment and can serve as an example for building a context-specific validity argument for the interpretations and uses of other language assessments designed for YLLs.

Perhaps most importantly, the body of validity evidence on the TOEFL Junior tests presented here can help and guide score users and other stakeholders to make informed decisions about the intended uses and interpretations of TOEFL Junior test scores and their fitness for different educational contexts. Nonetheless, given that test validation is an ongoing process, subsequent investigations or replications of existing studies in different linguistic, cultural, and proficiency contexts will continue to be informative and critical to enhance the TOEFL Junior validity argument.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, *4*(3), 3–30.

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

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

Bailey, A. L. (2007). Introduction: Teaching and assessing students learning English in school. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 1–26). Yale University Press.

Bailey, A. L., & Heritage, H. M. (2008). *Formative assessment for literacy, Grades K–6: Building reading and academic language skills across the curriculum*. Corwin Press.

Bailey, A. L., & Heritage, H. M. (2014). The role of language learning progressions in improved instruction and assessment of English langauge Learners. *TESOL Quarterly*, *48*(3), 480–506. https://doi.org/10.1002/tesq.176

Baron, P. A., & Papageorgiou, S. (2016). *Setting language proficiency score requirements for English-as-a-second-language placement decisions in secondary education* (Research Report No. RR-16-17). ETS. https://doi.org/10.1002/ets2.12102

Baron, P. A., & Tannenbaum, R. J. (2011). *Mapping the TOEFL Junior® test onto the Common European Framework of Reference* (Research Report No. RM-11-07). ETS. http://www.ets.org/Media/Research/pdf/RM-11-07.pdf

Brunfaut, T., Michel, M., & Ratajczak, M. (2021). Testing young foreign language learners' reading comprehension: Exploring the effects of working memory, grade level, and reading task. *Language Testing*, *38*(3), 356–377. https://doi.org/10.1177/0265532221991480

Buck, G. (2001). *Assessing listening*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732959

Butler, Y. G. (2017). The role affect in intraindividual variability in task performance for young learners. *TESOL Quarterly*, *51*(3), 728–737. https://doi.org/10.1002/tesq.385

Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Language *Testing*, *27*(4), 443–469. https://doi.org/10.1177/0265532210367633

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

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The Companion to Language Assessment: Volume III. Evaluation, methodology, and interdisciplinary themes*. John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla110

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

Choi, J. S., & Loewen, S. (2022). Exploring young learners' strategic behaviors in a speaking test. *TESOL Quarterly*, *56*(4), 1384–1396. https://doi.org/10.1002/tesq.3136

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

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

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL®. *Assessing Writing*, *10*(1), 5–43. https://doi.org/10.1016/j.asw.2005.02.001

Eberharter, K., Kormos, J., Guggenbichler, E., Ebner, V. S., Suzuki, S., Moser-Frötscher, D., Konrad, E., & Kremmel, B. (2023). Investigating the impact of self-pacing on the L2 listening performance of young learner candidates with differing L1 literacy skills. *Language Testing*, *40*(4), 960–983. https://doi.org/10.1177/02655322221149642

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive EAP training and band score gains on IELTS. *IELTS Research Reports*, *4*, 207–254.

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

ETS. (2018). *Handbook for the TOEFL Junior® Tests*.

ETS. (2022a). *TOEFL Junior® Speaking Scoring Guide*. https://www.ets.org/pdfs/toefl/toefl-junior-speaking-scoring-guide.pdf

ETS. (2022b). *TOEFL Junior® Writing Scoring Guide*. https://www.ets.org/pdfs/toefl/toefl-junior-writing-scoring-guide.pdf

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

ETS. (2023). TOEFL Junior® Test Taker Handbook.

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist*, *35*(1), 39–50. https://doi.org/10.1207/S15326985EP3501_5

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

Frost, K., Elder, C., & Wigglesword, G. (2011). Investigating the validity of na integrated listening-speaking task: A discourse-analysis of test takers' oral performance. *Langauge Testing*, *29*(3), 345–369. https://doi.org/10.1177/0265532211424479

Galikyan, I., Madyarov, I., & Gasparyan, R. (2019). *Student test takers' and teachers' perceptions of the TOEFL Junior® Standard test* (Research Report No. RR-19-29). ETS. https://doi.org/10.1002/ets2.12264

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*(2), 123–145. https://doi.org/10.1016/S1060-3743(00)00019-9

Gu, L. (2015). Language ability of young English language learners: Definition, configuration, and implications. *Language Testing*, *32*(1), 21–38. https://doi.org/10.1177/0265532214542670

Gu, L., & Hsieh, C.-N. (2019). Distinguishing features of young English language learners' oral performance. *Language Assessment Quarterly*, *16*(2), 180–195. https://doi.org/10.1080/15434303.2019.1605518

Gu, L., Lockwood, J. R., & Powers, D. E. (2015). *Evaluating the TOEFL Junior® Standard Test as a measure of progress for young English language learners* (Research Report No. RR-15-22). ETS. https://doi.org/10.1002/ets2.12064

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), 399–420. https://doi.org/10.1080/15434303.2023.2288253

Hsieh, C.-N. (2023). The role of task types and reading proficiency on young Enlgish as a foreign language learners' writing performances. *TESOL Quarterly*. Advance online publication. https://doi.org/10.1002/tesq.3286

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.

Hsieh, C.-N., & Wang, Y. (2019). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing*, *36*(1), 27–50. https://doi.org/10.1177/0265532217734240

Huang, B. H., Bailey, A. L., Sass, D. A., & Chang, Y.-h. S. (2021). An investigation of the validity of a speaking assessment for adolescent English language learners. *Language Testing*, *38*(3), 401–428. https://doi.org/10.1177/0265532220925731

Hulstijn, J. H. (2015). Language proficiency in native and non-native speakers: Theory and research. *John Benjamins.* https://doi.org/10.1075/lllt.41

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Kamiya, N. (2017). Can the National Center Test in Japan be replaced by commercially available private English tests of four skills? In the case of TOEFL Junior® Comprehensive. *Language Testing in Asia*, *7*, Article 15. 10.1186/s40468-017-0046-z

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. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice. 23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

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

Kim, M. (2023). Exploring literal and inferential reading comprehension among L2 adolescent learners: The roles of working memory capacity, syllogistic inference, and L2 linguistic knowledge. *Reading and Writing*, *36*, 1085–1110. https://doi.org/10.1007/s11145-022-10320-3

Kim, M., Nam, Y., & Crossley, S. A. (2022). Roles of working memory, syllogistic inferencing ability, and linguistic knowledge on second language listening comprehension for passages of different lengths. *Language Testing*, *39*(4), 593–617. https://doi.org/10.1177/02655322211060076

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

Kormos, J., Brunfaut, T., & Michel, M. (2020). Motivational factors in computer-administered integrated skills tasks: A study of young learners. *Language Assessment Quarterly*, *17*(1), 43–59. https://doi.org/10.1080/15434303.2019.1664551

Ling, G., & Gu, L. (2019). *Is greater access to English language learning associated with better performance on the TOEFL Junior® Comprehensive test? An exploratory investigation* (Research Report No. RR-19-17). ETS. 10.1002/ets2.12254

Madyarov, I., Movsisyan, V., Madoyan, H., Galikyan, I., & Gasparyan, R. (2021). *New validity evidence on the TOEFL Junior® Standard test as a measure of progress* (Research Report No. RR-21-19). ETS. 10.1002/ets2.12334

Manna, V., Yoo, H., & Monfils, L. (2018). *Evaluating invariance in test performance for adolescent learners of English as a foreign language* (Research Report No. RR-18-21). ETS. 10.1002/ets2.12208

Messick, S. (1996). Validity and wash back in langauge testing. *Langauge Tesitng*, *13*(3), 241–256. https://doi.org/10.1177/026553229601300302

Michel, M., Kormos, J., Brunfaut, T., & Ratajczak, M. (2019). The role of working memory in young second language learners' written performances. *Journal of Second Language Writing*, *45*, 31–45. https://doi.org/10.1016/j.jslw.2019.03.002

National Education Examinations Authority. (2018). *China's Standards of English Language Ability*. Higher Education Press.

National Governors Association & Council of Chief State School Officers. (2010). *Common Core state standards for English language arts & literacy in history/social studies, science and technical subjects*.

Nguyen, T. M. H., Gu, P., & Coxhead, A. (2023). Argument-based validaiton of Academic Collcation Tests. *Language Testing. Advance online publication*. https://doi.org/10.1177/02655322231198499

Papageorgiou, S., & Cho, Y. (2014). An investigation of the use of TOEFL Junior® Standard scores for ESL placement decisions in secondary education. *Language Testing*, *31*(2), 223–239. https://doi.org/10.1177/0265532213499750

Papageorgiou, S., Morgan, R., & Becker, V. (2015). Enhancing the interpretability of the overall results of an international test of English-language proficiency. *International Journal of Testing*, *15*(4), 310–336. https://doi.org/10.1080/15305058.2015.1078335

Papageorgiou, S., Wu, S., Hsieh, C.-N., Tannenbaum, R. J., & Cheng, M. (2022). Aligning language test scores to local proficiency levels: T le case of China's Standards of English Language Ability (CSE). *Chinese/English Journal of Educational Measurement and Evaluation/教育测量与评估双语季刊*, *3*(1), Article 1. 10.59863/CIPH5850

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly*, *12*(2), 153–177. https://doi.org/10.1080/15434303.2015.1008480

Schmidgall, J. (2017). *Articulating and evaluating validity arguments for the TOEIC® Tests* (Research Report No. RR-17-51). ETS. 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. 10.1002/j.2333-8504.2010.tb02235.x

So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale international English test. *Language Assessment Quarterly*, *11*(3), 283–303. https://doi.org/10.1080/15434303.2014.936936

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). *TOEFL Junior® design framework* (Research Report No. RR-15-13). ETS. 10.1002/ets2.12058

Swain, M., Huang, L.-S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT (SSTiBT): Test-takers' reported strategic behaviors*. (Research Report No. RR-09-30). ETS. 10.1002/j.2333-8504.2009.tb02187.x

Tannenbaum, R. J., & Baron, P. (2015). *Mapping scores from the TOEFL Junior® Comprehensive test onto the Common European Framework of Reference (CEFR)*. (Research Report No. RM-15-13). ETS. https://www.ets.org/Media/Research/pdf/RM-15-13.pdf

Timpe-Laughlin, V. (2018). *A good fit? Examining the alignment between the TOEFL Junior® Standard Test and the English as a foreign langauge currculum in* Berlin, Germany (Research Report No. RM-18-11). ETS. https://www.ets.org/Media/Research/pdf/RM-18-11.pdf

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

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

Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*, *72*(1), 5–44. https://doi.org/10.1111/lang.12424

Wallace, M. P. (2021). Exploring the relationship between L2 listening and metacognition after controlling for vocabulary knowledge. *Journal of Language and Education*, *7*(3), 187–200. 10.17323/jle.2021.12685

Wallace, M. P., & Lee, K. (2020). Examining second language listening, vocabulary, and executive functioning. *Frontiers in Psychology*, *11*, Article 1122. 10.3389/fpsyg.2020.01122

Wolf, M. K., Lopez, A. 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/10.4324/9781003274889-13

Wolf, M. K., Oh, S., Wang, Y., & Tsutagawa, F. S. (2018). Young adolescent EFL students' writing skills development: Insights from assessment data. *Language Assessment Quarterly*, *15*(4), 311–329. https://doi.org/10.1080/15434303.2018.1531868

Yeom, S., & Jun, H. (2020). Young Korean EFL learners' reading and test-taking strategies in a paper and a computer-based reading comprehension. *Language Assessment Quarterly*, *17*(3), 282–299. https://doi.org/10.1080/15434303.2020.1731753

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

Young, J. W., Morgan, R., Rybinski, P., Steinberg, J., & Wang, Y. (2013). *Assessing the test information function and differential item functioning for the TOEFL Junior® Standard test* (Research Report No. RR-13-17). ETS. 10.1002/j.2333-8504.2013.tb02324.x

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. ETS.

Suggested citation:

Hsieh, C.-N. (2024). *Building a validity argument for the TOEFL Junior® Tests* (TOEFL Research Report No. RR-102). ETS. https://doi.org/10.1002/ets2.12379

Action Editor: Larry Davis

Reviewers: Veronika Timpe-Laughlin and Mikyung Wolf