Measuring the Power of Learning.™

Research Memorandum
ETS RM–16-02

Designing the *TOEFL® Primary™* Tests

Yeonsuk Cho

Mitch Ginsburgh

Rick Morgan

Brad Moulder

Xiaoming Xi

Maurice Cogan Hauck

February 2016

ETS Research Memorandum Series

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

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

Designing the *TOEFL® Primary™* Tests

Yeonsuk Cho, Mitch Ginsburgh, Rick Morgan, Brad Moulder,
Xiaoming Xi, and Maurice Cogan Hauck

Educational Testing Service, Princeton, New Jersey

February 2016

Corresponding author: YCho@ets.org

Suggested citation: Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2016). *Designing the TOEFL® Primary™ tests* (Research Memorandum No. RM-16-02). Princeton, NJ: Educational Testing Service.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/

To obtain a copy of an ETS research report, please visit http://www.ets.org/research/contact.html

Action Editor: Donald Powers

Reviewers: Lin Gu and Spiros Papageorgiou

Abstract

The *TOEFL*® *Primary*™ tests were introduced in November 2013 for the standardized English proficiency assessment of young English-as-a-foreign-language (EFL) learners around the world. The tests are intended for children between approximately 8 and 12 years of age who are learning English in countries where English is not used for daily communication. The purpose of this document is to provide an overview of the development of the TOEFL Primary tests. The paper is organized into three parts. The first part presents the conceptual work that was undertaken to describe the intended population of test takers and to formulate a construct definition of young EFL learners' English ability. In the second part, we explain how the TOEFL Primary tests were designed. In the final part, we discuss the research needed to build validity arguments for the new tests and to advance our understanding of their intended population.

Key words: assessment, English as a foreign language (EFL), elementary school, test design, TOEFL Primary, young learners

Children around the world are learning English at much younger ages than ever before. A recent British Council survey of English teaching practices in 64 countries showed that, in many countries, children start to learn English before the age of 7 as part of their primary or elementary education (Rixon, 2013). To meet the assessment needs of this growing segment of the English as a foreign language (EFL) learner population, Educational Testing Service (ETS) developed the *TOEFL® Primary*™ tests as standardized English-language proficiency tests.

In this paper, we describe the process of developing the TOEFL Primary tests of reading, listening, and speaking from conceptualization to test design. We devote the first part of the paper to a discussion of the tests' purposes and intended uses and, based on a synthesis of research, to key aspects of language learning in children, all of which served as a conceptual guide in the test design. In the second part of the paper, we describe how the final design of the TOEFL Primary tests was shaped throughout the test development process. We conclude the paper with a discussion of the research that is needed to evaluate claims about the tests and improve their quality.

Intended Population

The TOEFL Primary tests are designed to serve children between 8 and 12 years of age[1] who are both learning English in countries where English is a foreign language and have limited opportunities to use English, either inside or outside the classroom (e.g., Carless, 2002; Kang, 2008; Munera, McNulty, & Ortiz, 2004; Tahira, 2012). We might call this group the general population of young EFL learners. These students typically learn English in classroom lessons taught by nonnative English-speaking teachers. They receive only a few English classes per week, perhaps only one, and instruction focuses on the development of communication skills in English (Inbar-Lourie & Shohamy, 2009; McKay, 2006).

There are also two other major types of EFL learning programs, according to McKay (2006), but they are less common. Some students may learn English in what is called language awareness programs. The instructional objective of these programs is to build awareness of another language rather than communicative competence. Students in language awareness programs naturally have much less exposure to English compared to those learning English in the classrooms described above. A small number of students may learn English in the other type of programs, called immersion, which are designed to optimize English language learning by teaching subjects such as science and history in English (Inbar-Lourie & Shohamy, 2009;

McKay, 2006). Due to differences in instructional goals and the type and amount of exposure to English, students in these less typical programs are likely to have different levels of English-language proficiency than the general population of young EFL learners described earlier. Because young EFL students in language awareness or immersion programs are relatively uncommon, they are generally not the main population for which the TOEFL Primary tests have been created.

The students in the TOEFL Primary tests' intended population possess a range of levels of English-language proficiency. The Council of Europe (n.d.) has suggested that achieving proficiency levels from A1 to B1 on the Common European Framework of Reference (CEFR) is a reasonable expectation for most elementary or low secondary students. The same targeted proficiency range was reported for primary or elementary school students in 21 of the countries in Rixon's (2013) survey of EFL teaching practices, with A2 being the most commonly targeted level. However, the range of proficiency levels in the TOEFL Primary population may be wider because opportunities to learn English outside the classroom vary from student to student. Young EFL students may attend private English-language enrichment classes or have tutors for additional support. This aid helps some students acquire more advanced English knowledge and skills than are expected of students whose only English study occurs in regular school-based English classes at the primary and elementary school levels, as reported in some studies (e.g., Chien, 2012; Garton, Copland, & Burns, 2011). Thus, the TOEFL Primary tests are developed to cover a wide range of proficiency levels represented among young EFL learners in typical EFL classrooms.

Test Purposes and Intended Uses

The TOEFL Primary tests measure general English-language proficiency. As independent measures of English-language proficiency, the TOEFL Primary tests of reading, listening, and speaking each offer users a snapshot of English ability in a particular skill area. To describe the intended uses and impacts of TOEFL Primary test scores succinctly, we have developed a theory of action (shown in Figure 1) that visually represents causal relationships between the components of the TOEFL Primary tests and their intended effects. Bennett (2010) suggested that the use of a theory of action helps articulate the intended effects of an educational assessment and identify areas of the research needed to evaluate its claims.

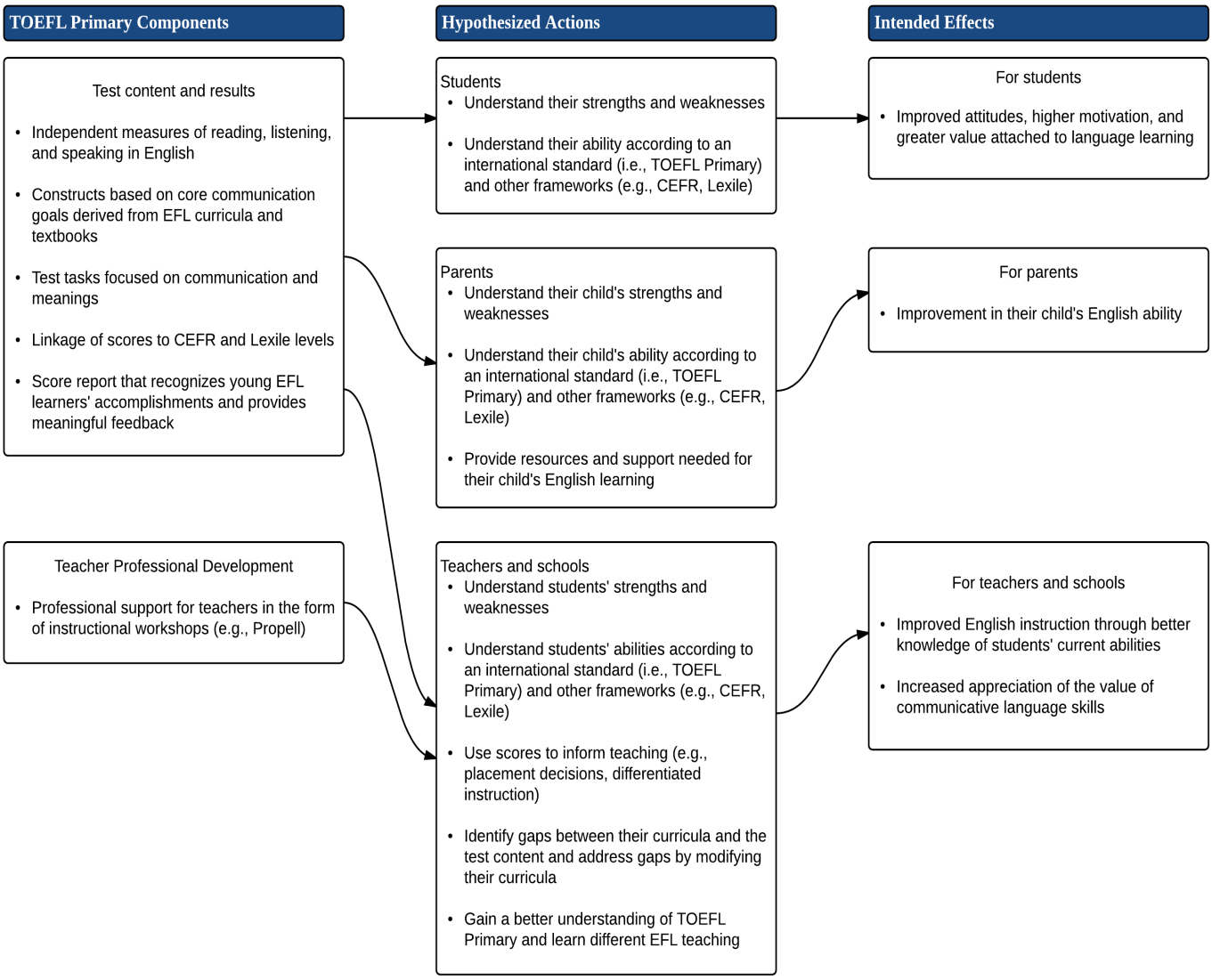

Figure 1. Theory of action for the TOEFL Primary.

The TOEFL Primary tests are intended primarily to support teaching and learning by providing meaningful feedback that educators can incorporate into their instruction. We expect that educators and parents will use TOEFL Primary scores to determine what their students and children have accomplished and to identify areas for improvement. The TOEFL Primary tests measure core communication skills derived from EFL curricula and provide detailed and meaningful feedback, and we expect that teachers and parents should find the test results helpful in providing instructional support that is appropriate to their learners' ability levels. In addition, some schools may use TOEFL Primary scores to place students into levels of instruction, if appropriate. These actions are expected to enhance young EFL students' learning experiences and ultimately lead to improved English proficiency. It is not desirable to use TOEFL Primary test scores for high-stakes decisions, such as admitting students, evaluating teachers, or comparing or ranking individual students.

Conceptual and Practical Bases for Defining the English Ability of Young English as Foreign Language Learners

To understand how young EFL students learn English, and subsequently to define a construct of English ability, we considered (a) general insights from relevant literature and (b) the content of EFL curricula and textbooks. In this section, we discuss the key points related to the language learning of young EFL students, and we describe how information from EFL curricula and textbooks informed the construct definition of English ability for the TOEFL Primary tests.

Insights From Research

What characteristics are associated with the English language proficiency of young EFL learners? According to MacNamara (1972), children possess an innate desire to make sense of the world, and this desire also applies in learning a language. As MacNamara described, language spoken to young learners in their first language usually contains concrete information that they can understand without knowing language rules. During the first few years of language learning, children focus mainly on meaning for communication. Throughout this process, they acquire language rules subconsciously, almost as a byproduct of recognizing meaning. The acquisition of language rules is, thus, incidental for the first several years of first-language learning. MacNamara supported this position with examples of children's language performance.

He stated that children attend mainly to meaning, disregarding function words such as prepositions while still conveying the substance of their messages.

MacNamara's explanation has an intuitive appeal, and it may be relative to the experience of elementary school EFL students who often learn English without receiving explicit instruction on language rules. Others, including McKay (2006) and Cameron (2001, 2003), have also suggested on the basis of both theoretical and empirical evidence that children seek meaning first. On this basis, Cameron recommended that in order to optimize children's language learning, language tasks used in classrooms should be meaning focused. Rather than manipulating vocabulary and grammatical structure to understand and express ideas, young language learners rely on memorized "chunks" of language: strings of words that learners memorize without understanding their grammatical structure. This may explain why EFL classroom language tasks for children tend to focus on formulaic expressions in meaningful contexts (see, for example, Kang, 2008). Decontextualized tasks that require metalinguistic knowledge or explicit knowledge of language rules should be avoided with young learners (Pinter, 2006). McKay (2006) also advised against teaching explicit grammar rules to young learners.

With regard to assessing young learners, therefore, the elements of linguistic knowledge, such as grammar, vocabulary, and pronunciation, should be viewed as ancillary rather than central to the definition of language ability. Some evidence indicates that it is unrealistic to expect a steady, predictable increase in either grammatical accuracy or vocabulary knowledge for young EFL learners. Zangl (2000) explained that the language development of young EFL learners is an uneven progression consisting of "peaks and troughs" (p. 256). Students make progress through "seemingly regressive periods" (p. 256) in which they produce errors that they did not produce before. This progress is partly due to a shift from relying on memorized language chunks to manipulating language structures based on their understanding of rules. Zangl advocated for the importance of considering the different stages of foreign language development when evaluating young EFL learners and interpreting their test results.

Related to Zangl's (2000) point, Johnstone (2000) expressed doubt about the validity of language proficiency levels defined by existing scales and standards because of a lack of empirical evidence regarding the development of foreign language ability among young students. Although increasing grammatical accuracy and vocabulary are often viewed as indicators of

higher proficiency, empirical evidence suggests otherwise. For example, Johnstone described interview data from students between 5 and 12 years of age who were learning a foreign language. Although older students produced more language, their language showed little improvement over their young counterparts in terms of grammatical accuracy and size of vocabulary. Johnstone also described the students' language as consisting largely of memorized expressions, suggesting little ability to manipulate language structures. This observation is consistent with a study by Traphagan (1997), who observed elementary school children learning Japanese as a foreign language. Traphagan noted that young learners' use of a certain grammatical particle in Japanese exhibited fewer errors and pauses compared to adult learners of Japanese as a foreign language. She posited that this difference between adults and children may be due to children's tendency to learn a foreign language through memorized chunks, whereas adult and proficient foreign-language learners are more likely to go beyond the regurgitation of formulaic chunks and show ability to manipulate language structures. Traphagan (1997) also observed that more proficient learners tended to produce more words in the target language, similar to Johnstone's (2000) finding. Moreover, Traphagan (1997) found that the language produced by less proficient students generally consisted of short and simple responses.

Previous research also supports the notion that oral language, rather than written language, is central to the language ability of young EFL learners. This is reflected in the EFL curricula of many countries (Butler, 2009; Hasselgren, 2000; Pinter, 2006; Zangl, 2000), in which more emphasis is placed on oral skills during early foreign-language education. McKay (2006) noted that the content of most EFL instruction in the first 2 years focuses on listening and speaking, with reading and writing being taught later. McKay emphasized the importance of oral language work for foreign-language development, arguing that oral language should be a key component of a language assessment designed for young learners: "Oral language makes up the core of young language learners' curriculum. Hence, to skip over oral language and to assess language learning through reading and writing is to deny the essence of young learners' language learning" (p. 177).

A similar viewpoint was expressed earlier by Cameron (2001, 2003), who suggested that classroom activities for young EFL learners should center on fostering oral language skills. She further argued that the development of oral language skills supports learning about language use contexts and discourse features in English, that is, how ideas are connected in various types of

text (e.g., a conversation or a story) associated with different types of discourse (e.g., description or narration).

In summary, although the literature on young EFL learners is limited, it provides insight into how young students learn a foreign language and suggests important implications for the design of an EFL assessment for children. Examples of such implications include the desirability of using meaning-focused assessment tasks and avoiding overemphasis on grammatical accuracy and vocabulary size. If measures of grammatical accuracy and lexical knowledge are included as part of the language construct for young learners, expectations should be lower than they would be for older learners. Finally, any assessment for learners should reflect that oral language is a core component of young EFL learners' language ability.

Insights From Practice

What do young EFL learners learn, and what are they expected to do with English?

To gain a concrete understanding of how communication in English is defined and how it is fostered during the early years of EFL instruction, we conducted analyses of national primary (elementary) school English curricula and textbooks from Brazil, Chile, China, Egypt, Japan, Korea, the Philippines, Qatar, and Singapore (Turkan & Adler, 2011). These countries were chosen for their representation of major geographical regions and for the ease with which the researchers could access their EFL curricula and textbooks. We also consulted European Language Portfolios (ELPs) from various European countries to capture language standards for primary students. Primary ELPs include age-appropriate language performance descriptors associated with proficiency levels of the CEFR.

In many countries, the focus of EFL education policy for young learners is on developing the ability to communicate in English (e.g., Gorsuch, 2000; Li, 1998; Mikio, 2008; Wu, 2001). EFL curricula delineate what young EFL learners should learn and be able to do, typically in terms of language objectives, reflecting the educational vision and goals of the national agencies that create them. Language objectives are written with varying levels of specificity and granularity. They address many aspects of language use, including linguistic resources (e.g., grammar, vocabulary), language functions (e.g., comprehend, infer, explain), text type (e.g., description, narrative), and topics (e.g., people, animals, weather), thereby providing valuable evidence of what students are expected to learn.

In the development of the TOEFL family of assessments, we also investigated language use contexts to gain a concrete understanding of the types of language and language-based tasks that are typical in each assessment's target-language use domain. Because it is reasonable to assume that children in EFL contexts have little opportunity to use English outside of the classroom, we limited the target-language use domain for TOEFL Primary to the EFL classroom and attempted to understand how English is used in that domain. In this regard, EFL textbooks are a useful resource, providing specific evidence of how language objectives are taught and how they relate to language use activities.

Our analyses of EFL curricula and textbooks focused on English-language objectives and language tasks (Turkan & Adler, 2011). Turkan and Adler (2011) indicated that a great deal of commonality exists in language objectives across EFL curricula. According to their analyses, elementary EFL education generally focuses on the development of oral language (i.e., listening and speaking), reading skills, and occasionally on rudimentary elements of writing. Minimal explicit attention is given to language rules. Classroom language tasks are designed to provide learners with opportunities to use English for communication by engaging them in age-appropriate and meaning-focused activities. Language activities are typically organized around a theme (e.g., my weekend) so that students are able to use learned expressions in a variety of contextualized settings (e.g., plan a weekend with a classmate or survey the class on favorite weekend activities). The language use contexts replicated in the EFL classroom are largely social, meaning that learners use language primarily to communicate with people around them (e.g., family, friends, classmates, teachers) on familiar topics (e.g., people, animals, school) and to obtain basic information from familiar sources (e.g., stories, announcements, directions).

Informed by the above results, we developed a construct definition of English communication for young EFL learners between 8 and 12 years of age for three language skills: reading, listening, and speaking (Figures 2–4). The writing construct was not initially considered for TOEFL Primary tests, given that little emphasis is placed on this skill during the early years of EFL education.

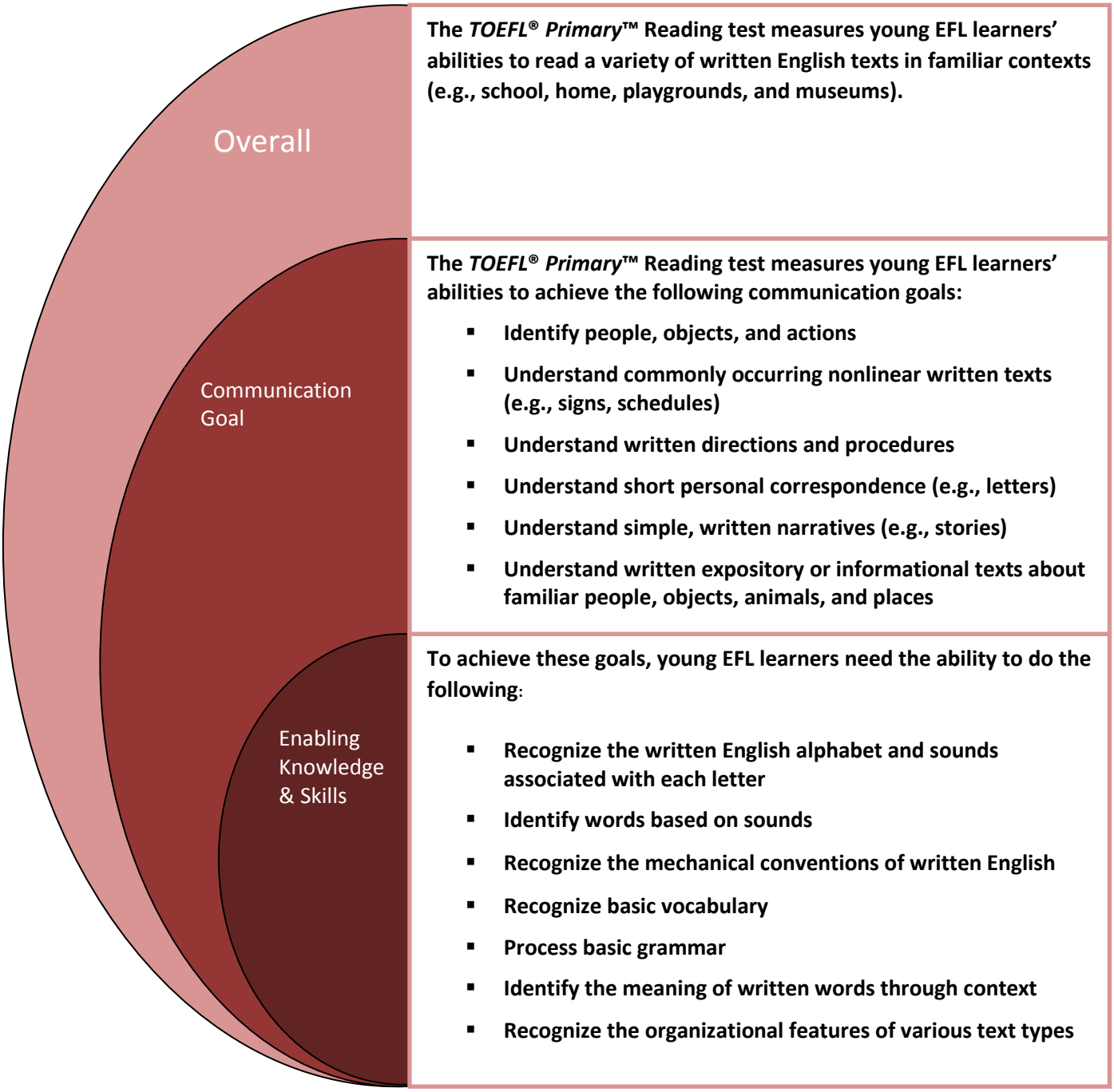

Figure 2. The construct of reading.

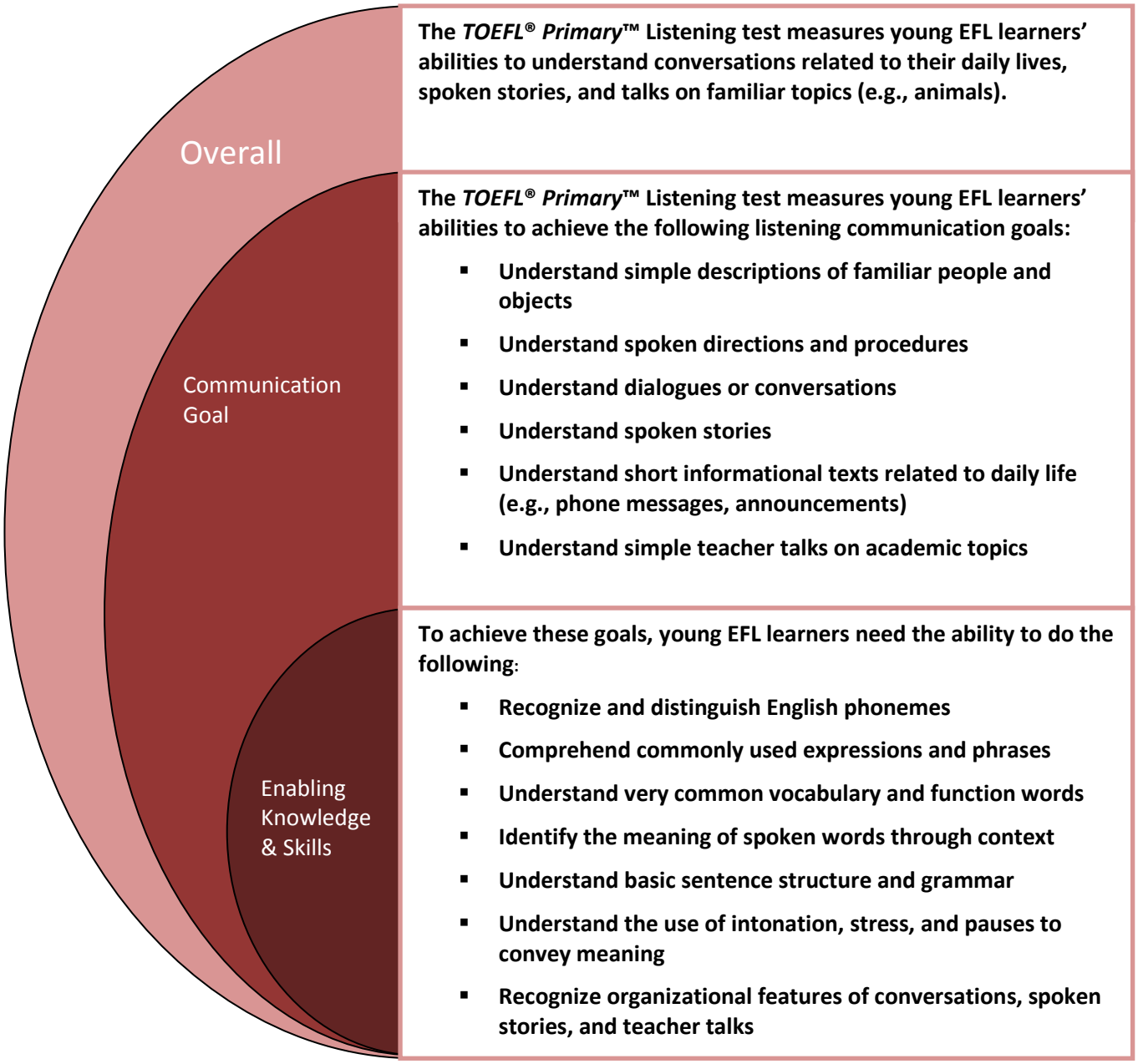

Figure 3. The construct of listening.

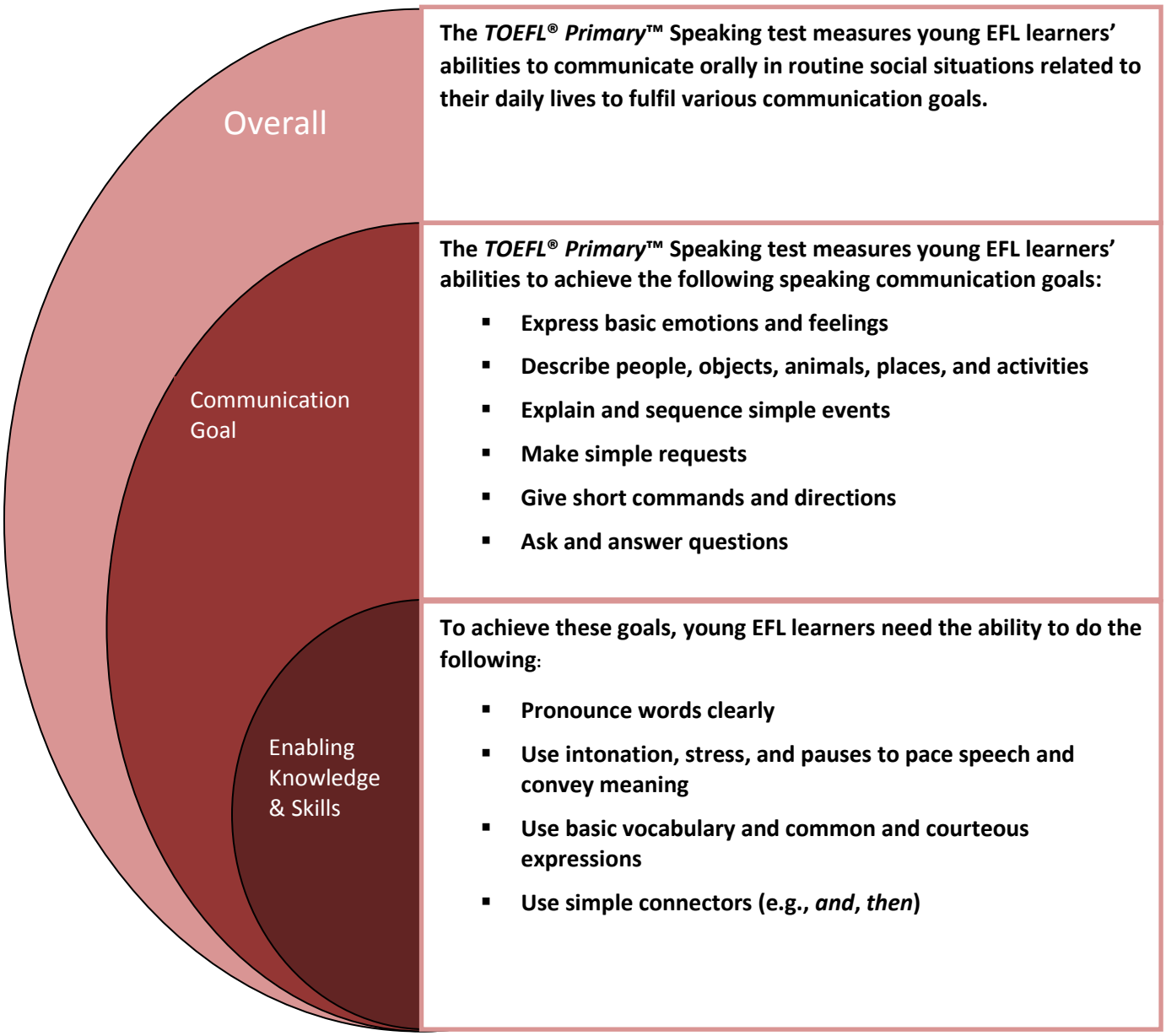

Figure 4. The construct of speaking.

For each skill, the TOEFL Primary construct definition first provides an overall, high-level definition of what it means for young EFL learners to use English for communication in each skill. Each overall definition is articulated in terms of communication goals and underlying language resources. Communication goals refer to types of communication that young EFL learners attempt in settings specified in the overall definition. The communication goals provide a basis for developing language test task types, which are shown later in the discussion of the test design. Figures 2 through 4 also present the language knowledge (e.g., vocabulary and grammar) and skills (e.g., identify characters) that young EFL students need to achieve communication goals. These linguistic resources are specified as enabling knowledge and skills, which are called on to accomplish communication goals.

These enabling knowledge and skills are part of the task design and may be measured directly if deemed appropriate in the context of assessing different levels of complexity within the communication goals. In particular, the use of matching pictures and words is justifiable with EFL students with less developed English abilities. These students are likely to rely on content words or phrases to communicate in English. Similarly, it is reasonable to use questions asking about how a particular word is used in the context of a reading passage in order to assess the reading abilities of advanced EFL learners.

Design of the TOEFL Primary

The purpose of this section is to present the content and structure of the TOEFL Primary tests. The final design is the result of a multistage test development effort: three rounds of prototyping followed by pilot testing and field testing. During these development stages, all aspects of the assessment were evaluated to ensure that the tasks and procedures would be relevant to and age appropriate for young test takers and that score results would achieve the desired psychometric qualities. A high-level description of the test development process is provided below to explain how information from various test development stages, as well as practical considerations, influenced the final test design.

Test Development Process and Key Design Considerations

A variety of assessment task types were proposed, based on the aforementioned construct definition, to measure the English ability of young EFL students. To evaluate the appropriateness of new item types and assessment procedures and to collect detailed feedback, prototyping studies were conducted with small groups of young English learners in four countries (China,

Korea, Mexico, and the United States) as well as with native English-speaking peers. In these studies, students were interviewed after trying the prototype test items. Concurrently, EFL teachers' evaluations of the new task types were gathered in China, Korea, and Mexico.

Results of the prototyping studies contributed to three main changes in our understanding of the construct and our approach to task design. One important change was to emphasize tasks measuring the ability to achieve communication goals over tasks measuring discrete language skills. Initially, a wide range of tasks types—from a measure of phonemic awareness to a measure of story comprehension—were prototyped to approximate language activities familiar to young EFL learners. For example, one set of prototype language tasks included a phonemic awareness task, which was designed to measure the ability to match sounds to letters. According to our analysis of curricula and textbooks, this ability appeared to be a common language objective across countries. Feedback from teachers, however, indicated that even though language tasks focusing on discrete knowledge and skills are relevant, they are not as valuable as communication-oriented tasks that focus directly on students' ability to use English to accomplish communication goals. This finding was also supported by data from cognitive interviews with students (for more detail, see Cho & So, 2014). As a result, tasks directly measuring enabling knowledge and skills are not used extensively on the TOEFL Primary tests.

Another major change that arose from findings of the prototyping studies concerned the administration format of speaking test items. Various administration formats were considered in the beginning. Speaking administration formats can be broadly divided into two categories: computer-administered (defined herein as a format in which test takers speak to a computer that records their responses) and direct face-to-face assessment (in which test takers speak to a human interlocutor). The face-to-face format can be divided further into individual and group (or pair) assessment, both being administered by an interviewer. With young students, face-to-face administration is generally thought to make speaking tasks more interactive and engaging. Thus, face-to-face individual and group administrations were initially attempted. These two administration models assumed that certified local teachers, not test takers' own teachers, would administer the test and score the responses. During the prototyping study, however, limitations of the face-to-face speaking assessment format became apparent. Some children felt uncomfortable talking to an adult they had never met. Other limitations included a lack of qualified teachers who could administer the test and evaluate the spoken responses and, in the group or pair administration model, unequal assessment opportunities among test takers. Given these findings,

a computer-administered assessment format was chosen. This decision, in turn, allowed test developers to explore a wider range of engaging contexts and innovative assessment types using technology. Further information about the design of the speaking tasks is presented later in this section.

Finally, through prototyping studies, a number of design issues, both general and specific, were identified for some of the reading and listening tasks, leading to the revision of test and item-writing specifications. Examples of general issues include unclear task descriptions, memory load, and the linguistic complexity of questions and response options. Based on feedback gathered in the prototyping studies, many task types underwent several iterations to mitigate potential sources of construct-irrelevant score variance (for further information, see Cho & So, 2014).

Pilot studies followed the prototyping studies to explore the influence of task features on item difficulty and to inform test form composition. Data from the pilot test administrations were collected from more than 1,300 students across eight countries for the reading and listening tests and from more than 400 students across 11 countries for the speaking test. The pilot data were analyzed to (a) obtain item parameters and reliabilities, (b) evaluate the relationship among item types and test sections, (c) determine optimal form composition to maximize reliability while minimizing the number of test items, and (d) decide a reading test length appropriate for test takers. On the basis of this analysis, item and test specifications were revised again. The analysis of pilot study results also led to two major test design decisions regarding the reading test length and the distribution of content on the reading and listening tests.

First, the length of the current reading test was informed by an analysis of speededness in performance data. Since children generally have shorter attention spans than adults (e.g., Robert, Borella, Fagot, Lecerf, & de Ribaupierre, 2009), it was decided during the conceptualization stage that the maximum testing time for each skill must not to exceed 30 minutes. In addition, the generally accepted desire not to overburden students with too much difficult content and reading load needed to be considered when deciding on test characteristics. Two reading test forms with different numbers of items and with common questions positioned early, middle, and late in the test forms were piloted to determine an appropriate test length. The results indicated that students' performance declined on the longer test form, thus supporting the use of fewer items.

The other important decision that emerged from the pilot study results was to offer the reading and listening tests at two difficulty levels, Step 1 and Step 2, while keeping the speaking test as a single-level test. Pilot testing showed performance differences in the reading and listening skills of learners among participating countries, suggesting that a single-level test could not effectively address the assessment needs of young EFL learners from different backgrounds. These results might be explainable by differences in English policy and curriculum among the participating countries. Two additional possible explanations for the performance differences among the countries in the pilot study were offered: (a) that the difficulty of reading and listening items reflects the intrinsic difficulty of a particular communication goal, and (b) that a multiple-choice format, in which reading and listening skills are measured, limits a range of responses. The pilot-phase analysis indicated that some task types were either too easy or too challenging for test takers closer to the endpoints of the ability range, thus reducing utility for measurement. For example, a task in which test takers match a single word to a picture is inherently simpler than one in which test takers answer comprehension questions after reading a long text. The picture-matching task provides little opportunity for proficient EFL students to demonstrate what they are able to do, whereas the reading comprehension task suffers from the same limitation for test takers with lower levels of proficiency. Given the results of the pilot study, and to make efficient use of testing time while maintaining sufficient test reliability, the test design was modified to include two levels of the reading and listening tests, with one level designed for students in the lower and middle range of proficiency assessed and one level designed for students in the middle and upper range. The item types used in each test level reflect concerns for both construct coverage and statistical characteristics.

Subsequent to the pilot study, large-scale field testing was conducted to obtain a sufficient number of items for test assembly and to develop a scale for score reporting. Data were obtained from more than 3,700 students across 14 countries for the reading and listening tests and from more than 500 students across 12 countries for the speaking test. Additional information about the design of the TOEFL Primary scoring system is described in the discussion of score reporting.

Test Content and Structure of the TOEFL Primary

On the basis of the pilot and field testing results, the final versions of the TOEFL Primary tests were assembled. Per the final test specifications, the TOEFL Primary Reading and

Listening tests are paper-based and consist of single-selection multiple-choice items. Task directions and items are presented in a color test book, and students mark their answers on a separate answer sheet. Listening test stimuli are played on audio CDs. The TOEFL Primary Speaking test is computer based and consists of constructed-response items. Students view and listen to test prompts on a computer and respond through microphones. Each test can be taken independently, depending on the child's English learning experience and assessment needs. The reading and listening tests are offered at two test levels; the speaking test is a single-level test.

Tables 1 and 2 show the blueprints of the two test levels for the TOEFL Primary Reading and Listening tests. Each test includes 30 multiple-choice questions that contribute to scores and a few pretesting questions that do not contribute to scores. The allotted time for a single test is 30 minutes for reading and approximately 30 minutes for listening, as the delivery of the student directions on the audio CD extends the testing time by a few minutes.

Each task type in the TOEFL Primary tests is designed to assess young EFL learners' ability to achieve one of the communication goals outlined in the test construct. In determining the composition of each test, input from experienced EFL teachers, in addition to performance data on each item type, was taken into consideration. During the pilot testing stage, 29 experienced EFL teachers from five countries were surveyed (Hsieh, 2013), and these teachers evaluated the degree of importance of individual communication goals measured by the TOEFL Primary tests and the effectiveness of the assessment task types.

In both the reading and listening tests, the majority of the communication goals are assessed at both Step 1 and Step 2, which explains the overlap in task types between the two steps (Tables 1 and 2). It should be noted, however, that although the same task types are shared for both steps, the level of linguistic complexity of stimuli in Step 1 is simpler than that in Step 2. Also, topics in Step 1 do not go beyond personal experiences and common, everyday surroundings.

Step 1 is suitable for EFL learners who can comprehend the following:

- basic formulaic expressions,
- basic vocabulary and phrases related to common objects and people,
- short and simple utterances related to survival needs (e.g., simple requests or directions), and
- short and simple texts relevant to students' daily lives (e.g., schedules or phone messages).

Table 1. Reading Test Structure

Item type	Task description	Communication goal	N[a] items Step 1[b]	N[a] items Step 2[b]
Match picture to word	Match a picture to one of three words.	Identify people, objects and actions.	6	
Match picture to sentence	Match a picture to one of three sentences.	Identify people, objects and actions.	7	
Sentence clue	Read a three-to-four-sentence description and identify what is being described.	Understand written expository or informational texts.	5	7
Telegraphic	Answer questions about a poster, a schedule, etc.	Understand nonlinear texts.	8	4
Correspondence	Read a letter or e-mail and answer comprehension questions.	Understand personal correspondence.	4	4
Instructional	Read directions and answer comprehension questions.	Understand written directions and procedures.		3
Narrative	Read a story and answer comprehension questions.	Understand narratives or stories.		8
Short expository	Read an expository text and answer comprehension questions.	Understand written expository or informational texts.		4
Total			30	30

[a]The number of items represents items that contribute to a test score. An operational test is longer because it includes pretesting items. [b]Shaded cells indicate that item types do not appear in the test level.

Table 2. Listening Test Structure

Item type	Task description	Communication goal	N[a] items Step 1[b]	N[a] items Step 2[b]
Listen and match	Listen to a sentence and select a corresponding picture.	Understand spoken descriptions.	7	
Follow directions	Listen to directions and select a corresponding picture.	Understand spoken directions and procedures.	7	6
Question-response	Listen to three versions of a two-turn conversation and choose a conversation that makes sense.	Understand dialogues or conversations.	6	
Dialogue	Listen to a dialogue and answer a comprehension question.	Understand dialogues or conversations.	5	5
Social-navigational monologue	Listen to a phone message/announcement and answer a comprehension question.	Understand short spoken informational texts.	5	5
Narrative	Listen to a story and answer comprehension questions.	Understand spoken stories.		8
Academic monologue	Listen to an academic monologue and answer comprehension questions.	Understand simple teacher talks.		6
Total			30	30

[a]The number of items represents items that contribute to a test score. An operational test is longer because it includes pretesting items. [b]Shaded cells indicate that item types do not appear in the test level.

Step 2 is recommended for EFL learners who have the same skills listed above and can also comprehend the following:

- simple and short stories and conversations on topics beyond personal experiences,
- simple explanations of objects related to content learning, and
- unfamiliar words, given a sufficient amount of contextual clues.

The speaking test is a single-level test consisting of seven constructed-response items. The test is computer-administered and lasts about 20 minutes. Because of the use of open-ended item types, the speaking test can be taken by both Step 1 and Step 2 test takers. Similar to the reading and listening task types, the speaking task types are associated with the communication goals outlined in the test construct (see Table 3).

A unique design feature of the speaking test is the use of a context and fictional characters. For example, one of the speaking test forms uses a trip to the zoo as a context, consisting of a series of events upon which the speaking tasks are based. Throughout the test, virtual characters (e.g., a zookeeper, children) appear at different times, functioning as interlocutors.[2] This contextualization is intended to create an authentic communication purpose for each speaking task so that test takers feel engaged and motivated to respond to the speaking prompts.

Table 3. Speaking Test Structure

Item type	Task description	Communication goal	Maximum score points
Warm-up	Answer a question with one word.	Warm-up	Not scored
Expression	Express emotion or opinion in response to a question.	Express basic emotions, feelings, and opinions.	3
Description	Describe a picture.	Give simple descriptions.	3
Directions	Explain a procedure based on sequenced pictures or animations.	Give directions.	5 (or 10)[a]
Narration	Explain a sequence of events shown in pictures or animations.	Explain and sequence simple events.	5 (or 10)[a]
Questions	Ask three questions.	Ask questions.	3
Requests	Make requests.	Make requests.	3
Total			27

[a]Each speaking test form includes one additional direction or narration task.

A second unique design feature is the inclusion of fun elements and scaffolding to enhance children's test-taking experience. The fun elements, consisting of animations, playful characters, and whimsical content are used to keep test takers engaged and to elicit more spontaneous and natural responses. The speaking test also incorporates the concept of scaffolding into task design by providing relevant key words to test takers in advance or directing their attention to a particular aspect of a stimulus that test takers need to describe.

Reporting TOEFL Primary Scores

TOEFL Primary score reports provide numeric scores and written performance descriptors to aid score interpretation. In this section, we describe how responses are scored and how TOEFL Primary score levels were established.

Scoring TOEFL Primary Responses

The responses to the TOEFL Primary Listening and Reading tests are dichotomously scored as correct or incorrect. The total raw scores of correct responses from each test are then calibrated on a scale from 100 to 115. The development of this scale is discussed below. Responses to the TOEFL Primary Speaking test are scored by human raters at ETS, based on holistic scoring rubrics with either a 0- to 3-point scale or a 0- to 5-point scale, depending on the task type. Both scoring rubrics (see Appendices A and B) were developed based on performance data collected in the pilot and field test administrations. For most of the speaking item types, the responses were relatively short and could be adequately differentiated using a 0- to 3-point scale. More complex tasks (e.g., giving directions, explaining a sequence of events) elicited a broader range of performances, allowing for the development of a 0- to 5-point scale. In developing the scoring rubrics, the core scoring criterion was the degree to which the young EFL learners could effectively achieve the intended communication goal for a task. The rubrics identify three major dimensions that are taken into consideration in evaluating spoken responses holistically—language use, content, and delivery—with each dimension considered in relation to the clarity of overall meaning.

Developing Score Scales and Score Bands With Performance Descriptors

Following the field test administrations, we analyzed the reading and listening score data using a two-parameter logistic (2PL) item response theory (IRT). The field test yielded a usable

sample size of at least 1,200 for each multiple-choice item. The 2PL model relates the probability of responding correctly, given a person's ability level, to the difficulty of a test item.

Items from the overlapping field test forms were calibrated concurrently using ETS proprietary software, which uses a marginal maximum likelihood estimation method to estimate item parameters, including item difficulty and item discrimination. Item parameters for reading and listening were estimated and evaluated separately. Following the field test administrations, items with extreme difficulty or low discrimination were removed from the item pool. Tests were assembled that met both content and statistical specifications for each of the two test levels. Some overlap in test items between the two levels was allowed. Raw-to-ability scale conversions were created to report test performance on a single continuous scale across Step 1 and Step 2. Both the reading and listening tests have a separate scaled scores range of 100 to 115, with a scaled score of 100 indicating the lowest performance on both Step 1 and Step 2. Scaled scores range from 101 to 109 for Step 1 and from 104 to 115 for Step 2.

Unlike the reading and listening test results, the speaking test results are not converted to another scale prior to reporting. Instead, speaking results are reported as the sum of individual task ratings resulting from use of the tasks' rating scales. Total speaking scores range from 0 to 27. To ensure the consistency of speaking test scores, operational speaking forms with similar score means and standard deviations of scores were created. The average reliability of the test forms was .90, as computed from field test data.

In addition to the aforementioned numeric scores, TOEFL Primary score reports also provide a performance description with a band score. As discussed earlier, one of the main intended effects of TOEFL Primary is to encourage continued language learning efforts by individual students. As Roberts and Gierl (2010) noted, a score report "serves a critical function as the interface between the test developer and a diverse audience of test users" (p. 25). Ideally, score reports explain what is expected on the test and how test results should be interpreted. A numeric score by itself does not provide meaningful information that can be used to support teaching and learning.

To aid the interpretation of test results, score bands were derived to characterize a typical performance at a particular band or score range. Both the score bands and performance descriptions were developed through an item mapping procedure. Item mapping is one approach to relating content information to test scores so that scores convey meaningful information about test-taker performance (Tong & Kolen, 2010). In an item mapping procedure, both item

performance data and expert judgments are utilized to articulate the performance of typical test takers at a given performance level, thus classifying test takers into different levels of ability. For an optimal level of classification accuracy and consistency, based on analysis of the field test data, it was recommended that test performances be categorized into six bands for reading and listening (across Step 1 and Step 2) and five bands for speaking. According to Livingston and Lewis (1993), classification accuracy is the extent to which the classification of a test taker, based on observed test scores, corresponds to the classification of the test taker based on estimates of "true" score—the hypothetical score that test taker would receive if a test had no measurement error. Classification consistency measures the extent to which the classification of a test taker based on test scores is consistent across different test administrations.

Test developers reviewed the multiple-choice items, arranged in order of difficulty, to characterize performance levels. They analyzed the content of the items representing score bands and articulated the characteristics of the items in each band. For the speaking test, average score profiles were created based on total raw scores. Test developers reviewed typical speech samples corresponding to average score profiles and articulated typical characteristics of performance for each score band. During the content review, test developers observed that performance patterns across score bands reflected both (a) the difficulty levels of various communication goals and (b) the complexity of language and topics. For example, the item mapping data indicated that, among the target population, the ability to understand spoken narratives is more difficult to acquire than the ability to understand spoken directions. Performance descriptors capture these patterns and explain linguistic characteristics of individual items and types of language tasks represented at a given score range. The performance descriptors also include suggestions for improving students' English abilities, which are based on the characteristics of performance at the next higher score band. Performance descriptors were refined in an iterative process to present them in plain and descriptive language (without linguistic jargon) that parents and teachers can easily understand.

In addition to reporting TOEFL Primary scores, score reports also show how TOEFL Primary scores relate to scores or levels of other frameworks or references for language learning. For example, TOEFL Primary scores are linked to the Lexile framework so that test users can use Lexile scores to identify reading materials that match their current reading levels (Metametrics, 2013). TOEFL Primary scores are also linked to the CEFR for languages to help test users interpret scores in terms of a widely recognized proficiency scale. The linkage between

TOEFL Primary scores and these frameworks was established using the field testing data. Further information about the methodology or results of these linking studies can be found in reports by Cline, Sanford, and Aguirre (2011) and Baron and Papageorgiou (2014).

TOEFL Primary Research

At the start of test development, we proposed a research agenda to support test design decisions at each stage of test development and to identify areas of future research with the aim of building a validity argument for the TOEFL Primary tests, as has been done for the *TOEFL iBT®* test (Chapelle, 2008). Chapelle's validity argument framework consists of six types of inference: domain description, evaluation, generalization, explanation, extrapolation, and utility.

During test development, our efforts were largely intended to guide and support test design and, consequently, produced evidence related to limited types of inferences in the validity argument. During the conceptualization stage of the development of TOEFL Primary, we consulted multiple sources to define the targeted domain of language use and the test constructs (discussed earlier in the current paper) providing evidence related to the domain description inference. Multiple rounds of prototyping studies were conducted to evaluate the degree to which the proposed assessment tasks and procedures are appropriate for their measurement purpose, providing evidence to support the evaluation inference. Furthermore, evidence that the quality of the TOEFL Primary tests meets the psychometric standards for educational assessment, as demonstrated in the pilot and field studies, provides additional support for the evaluation inference. Finally, the research we conducted to support the development of score reports lends evidence for explanation and utility inferences.

Now that the TOEFL Primary tests are in operational use, research has expanded to gather evidence across all six of the aforementioned types of validity inference that constitute Chapelle's (2008) validity argument. For example, research is being conducted to evaluate the impact of the TOEFL Primary tests on teaching and learning for young EFL learners by analyzing educators' and parents' perceptions of English teaching and learning practices before and after the introduction of the TOEFL Primary tests. This type of research addresses a utility-related validity inference by evaluating our claim that TOEFL Primary supports teaching and learning. Another study being conducted as part of ongoing research seeks to improve our current domain specification by addressing young EFL learners' writing ability, which is currently excluded from TOEFL Primary's construct. These are just a few examples of an array

of research studies that are in progress. Based on ETS's long tradition of backing test-related claims with empirical evidence, we have established a robust research agenda, and we update our research topics continuously to monitor the quality of the new tests and to evaluate the validity of score uses and claims about the tests.

References

Baron, P. A., & Papageorgiou, S. (2014). *Mapping the TOEFL® Primary™ Test onto the Common European Framework of Reference* (Research Memorandum No. RM-14-05). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8,* 70–91.

Butler, Y. G. (2009). How do teachers observe and evaluate elementary school students' foreign language performance? A case study from South Korea. *TESOL Quarterly*, *43*(3), 417–444.

Cameron, L. (2001). *Teaching languages to young learners.* Cambridge, UK: Cambridge University Press.

Cameron, L. (2003). Challenges for ELT from the expansion in teaching children. *ELT Journal, 57*(2), 105–112.

Carless, D. (2002). Implementing task-based learning with young learners. *ELT Journal, 56*(4), 389–396.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). London, UK: Routledge.

Chien, C. W. (2012). Differentiated instruction in an elementary school EFL classroom. *TESOL Journal, 3*(2), 280–291.

Cho, Y., & So, Y. (2014). *Construct-irrelevant factors influencing young English as a foreign language (EFL) learners' perceptions of test task difficulty* (Research Memorandum No. RM-14-04). Princeton, NJ: Educational Testing Service.

Cline, F., Sanford, E., & Aguirre, A. (2011, June). *Linking TOEFL Junior Reading scores to the Lexile measure.* Poster presented at the 33rd Language Testing Research Colloquium (LTRC), Ann Arbor, MI.

Council of Europe. (n.d.). *ELP checklists for young learners: Some principles and proposals.* Retrieved from https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016804932bd

Garton, S., Copland, F., & Burns, A. (2011). *Investigating global practices in teaching English to young learners* (ELT Research Papers No. 11-01). Birmingham, UK: British Council.

Gorsuch, G. J. (2000). EFL education policies and educational cultures: Influences on teachers' approval of communicative activities. *TESOL Quarterly, 34*(4), 675–710.

Hasselgren, A. (2000). The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing, 17*(2), 261–277.

Hsieh, C. (2013, September). *Establishing domain representations for a large-scale language assessment for young EFL learners*. Paper presented at the 15th Midwest Association of Language Testers (MwALT) Annual Conference, East Lansing, MI.

Inbar-Lourie, O., & Shohamy, E. (2009). Assessing young language learners: What is the construct? In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 83–96). Berlin, Germany: Moulton de Gruyter.

Johnstone, R. (2000). Context-sensitive assessment of modern languages in primary (elementary) and early secondary education: Scotland and the European experience. *Language Testing, 17*(2), 123–143.

Kang, D.-M. (2008). The classroom language use of a Korean elementary school EFL teacher: Another look at TETE. *System, 36*(2), 214–226.

Li, D. (1998). It's always more difficult than you plan and imagine: Teachers' perceived difficulties in introducing the communicative approach in South Korea. *TESOL Quarterly, 32*(4), 677–703.

Livingston, S. A., & Lewis, C. (1993). *Estimating the consistency and accuracy of classifications based on test scores* (Research Report No. RR-93-48). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1993.tb01559.x

MacNamara, J. (1972). Cognitive basis of language learning in infants. *Psychological Review, 79*(1), 1–13.

McKay, P. (2006). *Assessing young language learners.* Cambridge, UK: Cambridge University Press.

Metametrics. (2013). *Linking the TOEFL Primary with the Lexile Framework.* Unpublished manuscript.

Mikio, S. (2008). Development of primary English education and teacher training in Korea. *Journal of Education for Teaching, 34*(4), 383–396.

Munera, I. C. C., McNulty, M., & Ortiz D. I. Q. (2004). Elementary English language instruction: Colombian teachers' classroom practices. *PROFILE: Issues in Teachers' Professional Development, 5*(1), 37–55.

Pinter, A. (2006). *Teaching young language learners.* Oxford, UK: Oxford University Press.

Rixon, S. (2013). *Survey of policy and practice in primary English language teaching worldwide.* Retrieved from https://www.teachingenglish.org.uk/article/british-council-survey-policy-practice-primary-english-language-teaching-worldwide

Robert, C., Borella, E., Fagot, D., Lecerf, T., & de Ribaupierre, A. (2009). Working memory and inhibitory control across life span: Intrusion errors in the reading span test. *Memory and Cognition, 37*(3), 336–345.

Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice, 29*(3), 25–38.

Tahira, M. (2012). Behind MEXT's new course of study guidelines. *The Language Teacher, 36*(3), 3–8.

Tong, Y., & Kolen, M. (2010). Scaling: An ITEMS module. *Educational Measurement: Issues and Practice, 29*(4), 39–48.

Traphagan, T. W. (1997). Interviews with Japanese FLES students: Descriptive analysis. *Foreign Language Annals, 30*(1), 98–110.

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

Wu, Y. (2001). English language teaching in China: Trends and challenges. *TESOL Quarterly, 35*(1), 191–194.

Zangl, R. (2000). Monitoring language skills in Austrian primary (elementary) schools: A case study. *Language Testing, 17*(2), 250–260.

Appendix A. TOEFL Primary Speaking Scoring Guide: 0-3 Point Rubric

For the following communication goals: **express basic emotions, feelings, and opinions**
give simple descriptions
make simple requests
ask questions

Score	Language use, content, and delivery descriptors
3	The test taker achieves the communication goal. **A typical response at the 3 level is characterized by the following.** • **The meaning is clear. Minor errors in grammar or word choice do not affect task achievement.** • **The response is accurate and complete, and the content is appropriate for the task.** • **Speech is intelligible, and the delivery is generally fluid. It requires minimal listener effort for comprehension.**
2	The test taker partially achieves the communication goal. **A typical response at the 2 level is characterized by the following.** • **The meaning is mostly clear. Some errors in grammar or word choice may interfere with task achievement.** • **The response is not fully accurate or complete, or the content is not fully appropriate for the task.** • **Speech is generally intelligible, but the delivery may be slow, choppy, or hesitant. It requires some listener effort for comprehension.**
1	The test taker attempts to achieve the communication goal. **A typical response at the 1 level is characterized by the following.** • **The meaning is obscured because of frequent errors in grammar and word choice.** • **The response is inaccurate or incomplete, or the content is inappropriate for the task.** • **Speech is mostly unintelligible or unsustained. It requires significant listener effort for comprehension.**
0	**The test taker does not attempt to achieve the communication goal OR the response contains no English OR the response is off topic and does not address the prompt.**

Appendix B. TOEFL Primary Speaking Scoring Guide: 0-5 Point Rubric

For the following communication goals: explain and sequence simple events
give directions

Score	Language use, content, and delivery descriptors
5	The test taker fully achieves the communication goal. **A typical response at the 5 level is characterized by the following.** • **The meaning is clear. Grammar and word choice are effectively used. Minor errors do not affect task achievement. Coherence may be assisted by use of connecting devices.** • **The response is full and complete. Events are described accurately and are easy to follow.** • **Speech is fluid with a fairly smooth, confident rate of delivery. It contains few errors in pronunciation and intonation. It requires little or no listener effort for comprehension.**
4	The test taker achieves the communication goal. **A typical response at the 4 level is characterized by the following.** • **The meaning is mostly clear. Some errors in grammar and word choice may be noticeable, but the task is still achieved. Use of connecting devices to link ideas may be limited.** • **The response is mostly complete. Descriptions contain minor lapses or inaccuracies, but the events can still be readily followed.** • **Speech is mostly fluid and sustained, though some hesitation and choppiness may occur. It contains minor errors in pronunciation and intonation. It requires minimal listener effort for comprehension.**
3	The test taker partially achieves the communication goal. **A typical response at the 3 level is characterized by the following.** • **The meaning is sometimes obscured. Errors in grammar and word choice are noticeable and limit task achievement. The response may include attempts to use connecting devices.** • **The response is somewhat complete. Lapses and inaccuracies require the listener to fill in the gaps between key events.** • **Speech may be sustained throughout, but the pace may be slow, choppy, or hesitant. It contains errors in pronunciation and intonation. It requires some listener effort for comprehension.**
2	The test taker is limited in achieving the communication goal. **A typical response at the 2 level is characterized by the following.** • **The meaning is often obscured because of errors in grammar and word choice. Attempts at using connecting devices are unsuccessful or absent.** • **The response is mostly incomplete. Multiple lapses and gaps make it difficult for listeners unfamiliar with the events to follow along. Meaningful content may be conveyed through repetition.** • **Speech is noticeably slow, choppy, or hesitant throughout and may include long pauses. It contains frequent errors in pronunciation and intonation. It requires listener effort for comprehension.**
1	The test taker attempts to achieve the communication goal. **A typical response at the 1 level is characterized by the following.** • **The meaning is obscured because of frequent errors. Grammar and word choice are extremely limited and often inaccurate.** • **The response is incomplete. Major lapses and gaps make events unclear. The response may consist of a single word or a few words related to the prompt. It may be highly repetitive.** • **Speech is not sustained or is mostly incomprehensible. It contains numerous errors in pronunciation and intonation. It requires significant listener effort for comprehension.**
0	**The test taker does not attempt to achieve the communication goal OR the response contains no English OR the response is off topic and does not address the prompt.**

Notes

[1] The TOEFL Primary tests can, however, be used with older students in certain educational contexts, if appropriate.

[2] For a sample TOEFL Primary Speaking test, see https://toeflprimary.caltesting.org/sampleQuestions/TOEFLPrimary/index.html.