Research Memorandum
ETS RM–20-08

Exploring the Alignment Between a Curriculum and a Test for Young Learners of English as a Foreign Language

Spiros Papageorgiou
Xiaoqiu Xu
Veronika Timpe-Laughlin
Deborah M. Dugdale

November 2020

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.

Exploring the Alignment Between a Curriculum and a Test for Young Learners of English as a Foreign Language

Spiros Papageorgiou[1], Xiaoqiu Xu[2], Veronika Timpe-Laughlin[1], and Deborah M. Dugdale[2]
[1]Educational Testing Service, Princeton, New Jersey, United States
[2]VIPKid, Beijing, China

November 2020

Corresponding author: S. Papageorgiou, E-mail: spapageorgiou@ets.org

Suggested citation: Papageorgiou, S., Xu, X., Timpe-Laughlin, V., & Dugdale, D. M. (2020). *Exploring the alignment between a curriculum and a test for young learners of English as a foreign language* (Research Memorandum No. RM-20-08). Educational Testing Service.

Action Editor: John Norris

Reviewers: Ching-Ni Hsieh and Mikyung Wolf

Abstract

The purpose of this study is to examine the appropriateness of using the *TOEFL Primary*® tests to evaluate the language abilities of students learning English as a foreign language (EFL) through an online-delivered curriculum, the VIPKid Major Course (MC). Data include student test scores on the TOEFL Primary Listening and Reading tests and expert judgments on the degree of alignment between the content of the tests and the learning activities included in VIPKid MC Levels 3–7. Analysis of the score data suggested that the TOEFL Primary Reading and Listening tests are, in general, at the appropriate level of difficulty for MC students. Additionally, the TOEFL Primary test score data indicated an increase in language proficiency across the VIPKid MC levels. The content of the TOEFL Primary tests and the learning activities included in the MC curriculum were found to be largely aligned with each other. We conclude with a discussion of the implications of this study, including (a) the use of EFL tests for young learners in general and (b) the use of the TOEFL Primary tests in the context of the VIPKid MC in particular.

Keywords: alignment, young learners, curriculum, assessment, *TOEFL Primary*®, VIPKid Major Course

Acknowledgments

We thank Amanda Mohan and Brandon Lambert for their help with data coding. We also thank Elaine Wu, Jacqueline Jarvis, and Jacqueline Chen for facilitating data collection and for reviewing earlier versions of the manuscript. Special thanks to our ETS colleagues John Norris, Mikyung Wolf, and Ching-Ni Hsieh for their careful review of an earlier version of the manuscript and their helpful comments.

Table of Contents

List of Figures

List of Tables

In the educational measurement literature, alignment typically refers to the extent to which the content of an assessment covers the skills and abilities described in K–12 *content standards*, which define what should be taught in the curriculum. *Performance standards*, on the other hand, define what learners must do in order to demonstrate performance at a given level (Hambleton, 2001) and are discussed later in this section. As Webb (2007) pointed out, the No Child Left Behind Act (2001) in the United States resulted in increased demand for assessments to demonstrate alignment with content standards in terms of comprehensiveness, emphasis, depth, and clarity for users. The Common Core State Standards (http://www.corestandards.org/resources), which describe the skills and abilities expected by students at each grade level, have further raised demand for aligning assessments to content standards. The demand for alignment of assessments to various content standards has also increased worldwide because of education reforms that push for accountability, including close monitoring of students' progress and use of standardized tests (Deville & Chalhoub-Deville 2011).

Although there is no single best approach for conducting an alignment study, Webb (2007) proposed four criteria to systematically evaluate the alignment of assessments to content standards. These criteria address questions related to content coverage and cognitive challenge for the students as follows:

- categorical occurrence: Does the test cover the content discussed in the standard?
- depth of knowledge consistency: Is the assessment as cognitively challenging for the test takers as might be expected in the standard?
- range of knowledge correspondence: How does the breadth of knowledge in the assessment correspond to the knowledge expected by the standard?
- balance of representation: How is specific knowledge emphasized in comparison to the standard?

In the language testing field, alignment has been associated with efforts to map (or link) test scores to external proficiency levels and descriptors, such as those in the Common European Framework of Reference (CEFR) of the Council of Europe (2001), which essentially

function as performance standards rather than content standards. Such alignment is typically seen as an approach to facilitate the interpretation of test scores (Tannenbaum & Cho, 2014). Through the mapping process, numeric scores are associated with the can-do statements of the CEFR, which describe what learners are able to do using a foreign language. The process of score mapping involves several interrelated steps, which examine both the alignment of the content of the test to the descriptions of the proficiency levels as well as how scores are mapped onto these levels (Council of Europe, 2009; Harsch & Hartig, 2015; Papageorgiou, 2016). Ultimately, alignment is a claim about the interpretation of test scores in relation to external levels of language proficiency. Data to support such a claim can range from experts' judgments in standard setting workshops to language proficiency test scores and other indicators of student performance, such as grade point average or teacher evaluations of student performance (Papageorgiou et al., 2015; Papageorgiou et al., 2019).

Although a large body of research has been published on the alignment of assessments to language proficiency levels and standards for more than a decade (see, e.g., papers in Figueras & Noijons, 2009; Martyniuk, 2010), little research has been published on the alignment of language assessments to English as a foreign language (EFL) curricula. Two such studies in the context of young learner (YL) assessment have investigated the fit of an external, large-scale assessment for a given local, educational context (e.g., Hsieh, 2015; Timpe-Laughlin, 2018). Hsieh (2015) examined the use and appropriateness of the *TOEFL Primary*® tests in an English-medium instructional context in Kenya. Hsieh reviewed the alignment between the content of the listening and reading tests of TOEFL Primary Step 2 and the Kenyan Standards (i.e., grade levels) 4 and 5. The author also analyzed data obtained from a TOEFL Primary test pilot in Kenya to examine test-taker performance on different task types. Additionally, the author conducted focus group interviews with Kenyan EFL teachers to gauge their perceptions about the appropriateness of using TOEFL Primary in their local teaching contexts. Hsieh found that, for example, pattern drills and vocabulary practice made up a large portion of the exercises whereas more communicative activities such as listening tasks were underrepresented. Moreover, although the difficulty of the various task types included in TOEFL Primary Step 2 corresponded to activities included in textbooks for Grades 4 and 5, textbooks

and instructional materials did not cover the entire range of text types on the TOEFL Primary, making it challenging for YLs to deal with these texts in the context of the assessment.

Timpe-Laughlin (2018) built upon Hsieh's (2015) line of research when examining the fit between the EFL curriculum mandated by the ministry of education in the state of Berlin, Germany, and the competencies and language skills assessed by the *TOEFL Junior*® test. She reviewed curricula and systematically coded activities for competencies and language skills in the textbooks of four grade levels. Additionally, Timpe-Laughlin interviewed teachers at different schools in order to gauge whether they regarded the external measure as an appropriate assessment for their learners. Although results suggested a good match between the contents covered in the four levels of EFL classes, findings also revealed areas in need of further research and development, such as including more diagnostic information on score reports.

The research strand described previously has been gaining ground insofar as it is important to select an assessment that is (a) age appropriate for YLs who are in the midst of rapid cognitive, social, and affective development (Barrouillet, 2015; Csapó & Nikolov, 2009; Nikolov & Timpe-Laughlin, 2020) and (b) adequately aligned to the curriculum, so that teachers can obtain sound information about their students' potential progress and ultimately help them achieve their learning goals (Herman & Webb, 2007; Roach et al., 2008). Both Hsieh (2015) and Timpe-Laughlin (2018) investigated YL contexts in which English was taught as a foreign language in traditional, face-to-face classroom contexts. To our knowledge, no alignment studies have examined the match between an external large-scale assessment for YLs and the content of an immersive EFL curriculum that is delivered remotely via video-mediated technology—a context that is quite distinct from the type of K–12 curricula explored in previous research. The context of our study is described next.

Context of the Study

The purpose of our study is to examine the appropriateness of using the TOEFL Primary tests, designed by Educational Testing Service (ETS), and in particular the test scores to evaluate the English language abilities of YLs in the context of the VIPKid Major Course (MC) Levels 3–7, an immersive EFL curriculum that is delivered online. Since 2016, VIPKid MC students have had

the option to take the TOEFL Primary tests so that teachers and parents can obtain information about their language ability. However, beyond a general review of the test content and the curriculum, no detailed investigation of whether the test is a good fit for that particular context has been conducted. Therefore, an important next step was to examine the fit more systematically, exploring in particular if test difficulty was appropriate for the VIPKid students and if the learning activities used in the VIPKid classes and the tasks in the TOEFL Primary test were aligned in terms of their content. Examining these two aspects is particularly important in the context of young English learners for the following reasons (see also Lee & Winke, 2018):

- When the gap between test task difficulty and student ability is too large, then student performance will not be measured reliably; thus, scores are unlikely to be useful.
- When the learning activities used in the classroom and the tasks in the test are aligned in terms of their content, students are likely to be familiar with test mechanics and their performance is less likely to be impacted by factors irrelevant to language ability, such as test-taking anxiety.

In this section we first describe the VIPKid MC and the TOEFL Primary tests, before presenting the research questions we sought to address.

VIPKid Major Course

VIPKid is a global online education technology company headquartered in Beijing, China, with its main U.S. office in San Francisco, California. The main course offered at VIPKid is the MC, in which EFL is taught in a one-on-one, online class format facilitated remotely by English teachers based in the United States or Canada. MC is VIPKid's flagship curriculum line and aims to provide English instruction for children in China ages 4–15. All VIPKid curricula, including the MC course, are designed to supplement what students learn in English classes in their formal schooling.

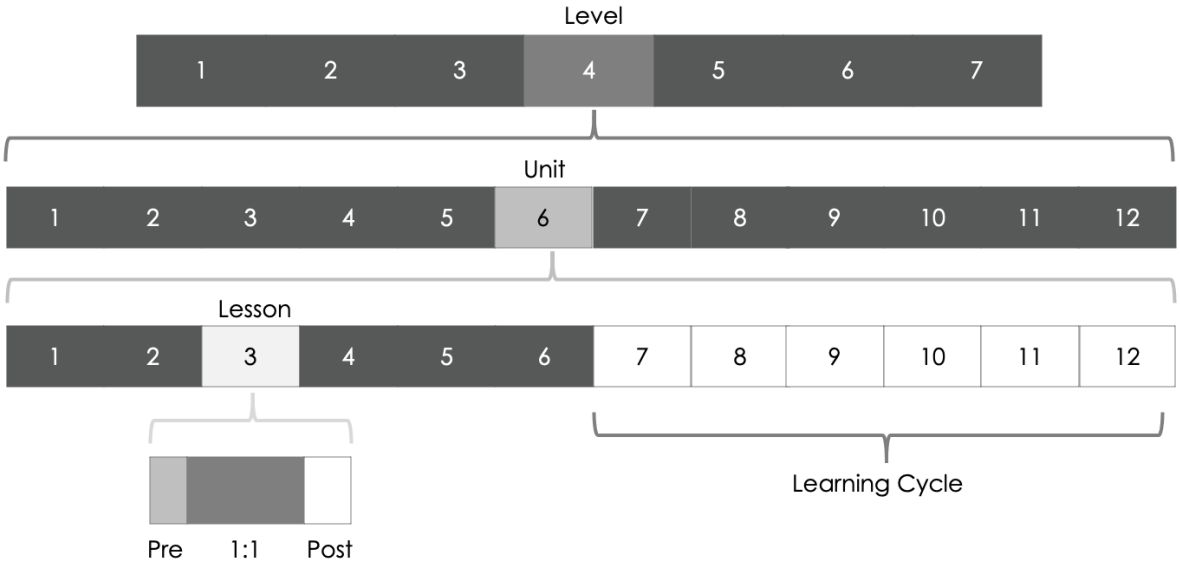
The VIPKid MC curriculum integrates Bloom's taxonomy (Krathwohl, 2002) and the concept of gradual release of responsibility (Fisher & Frey, 2013) in structuring lesson sequences. Additionally, it provides appropriate scaffolding for beginning students. As students

progress through the curriculum and develop fluency and increasing independence in language use, supports are gradually removed to allow students to successfully acquire and use their new language skills (Cameron, 2001; Samana, 2013). Accordingly, lessons in the MC curriculum include activities and tasks that match students' current proficiency level.

When our study was conducted, the MC curriculum progressed from entry-level English at Level 1 to academic English at Level 7. In the earliest levels of the curriculum (Levels 1–3), the students have their first experiences with the basic building blocks of English and are gradually introduced to grammatical structures. Partway through the curriculum (Levels 4–5), reading and expression are introduced to engage students with more abstract concepts such as genres and cultural and academic topics. At Level 6, emphasis is placed on reading skills to support academic studies in English. By the time students have completed Level 7, they are focused on building reading and writing skills as well as continuing to enhance their listening and speaking skills. It should be noted that Level 8 was added to the curriculum in Spring 2020, but it is not described in this report as our data collection was completed prior to the launch of Level 8.

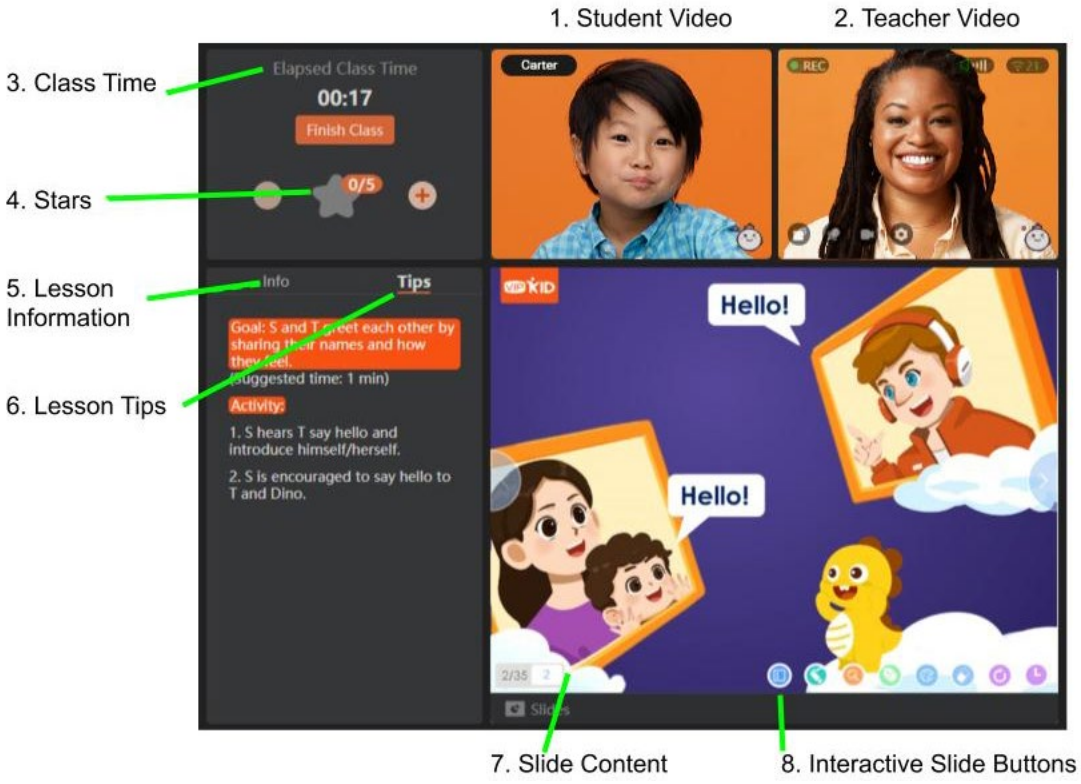
Figure 1 presents an example of the structure of the MC Level 4 curriculum. Levels 1–6 include 12 units each, whereas Level 7 includes six units. Units are thematically based. Each unit in Levels 2–6 consists of 12 lessons that are grouped into two 6-lesson learning cycles: one cycle for Lessons 1–6 and one cycle for Lessons 7–12. In Level 1, each unit consists of eight lessons grouped into four-lesson learning cycles. In Level 7, each unit has 24 lessons grouped into two 12-lesson learning cycles. For example, the larger theme in Level 4, Unit 6 is "Going Places" and the learning cycles are "In My Country" and "Traveling Abroad." A lesson comprises a preclass video, a 25-minute one-to-one class, and postclass enrichment activities.

Figure 1. Illustration of the Components of Level 4 in the VIPKids Major Course Curriculum



Students access their classes by logging into VIPKid's proprietary platform. They interact with materials in a "flipped classroom" style (Bishop & Verleger, 2013); that is, they are given the opportunity to preview content before class through videos, thus preparing for the in-depth engagement with their teacher during the 25-minute one-to-one class period. After class, they can further engage with the content by completing additional enrichment activities. Given that each class is being digitally recorded, students can also review classes in order to deepen their understanding of the class material. Teachers use VIPKid's proprietary platform to conduct classes. The platform includes lesson materials in the form of an interactive slide deck to guide the students' learning. The slides include learning activities and teaching tips for the teacher. Teachers follow the interactive slide decks for the 25-minute one-to-one class. An example of the lesson platform is shown in Figure 2, which is the teacher's view of an interactive slide.[1] While the core features of the platform (teacher and student video, and interactive slide) are visible to the student, additional features are only available to the teacher. The features of the platform in the teacher-view mode are shown in Figure 2.

Figure 2. Example of the Teacher-Side Lesson Platform



Note. Screenshot from VIPKid shows an example of a lesson platform. 1. Also available to the student. 2. Also available to the student. 3. Shows how much time is left. 4. Ranks participation in the class. 5. Pops up lesson objectives. 6. May pop up goals of the activity as well as actions for teacher and student. 7. Also accessible to the student. 8. Click to use actions to facilitate teaching, such as magnifying a part of the slide or writing or drawing on the slide.

The TOEFL Primary Tests

The TOEFL Primary tests are part of the *TOEFL*® family of assessments, offered by ETS. Along with the TOEFL Junior tests, they constitute the TOEFL Young Student Series (YSS), intended for YLs. The TOEFL Primary tests are designed for children between 8 and 12 years of age who learn English in countries where English is a foreign language. The tests measure young EFL students' abilities to communicate in English in three modalities: reading, listening, and speaking. The TOEFL Primary Reading and Listening test sections must be taken together as a single test. The TOEFL Primary Speaking test is taken as an independent test. TOEFL Primary

Reading and Listening tests are available at two difficulty levels (Step 1 and Step 2). The speaking test is a single-level test that both Step 1 and Step 2 test takers can take.

The TOEFL Primary tests focus on YLs' ability to use English in accomplishing communication goals in familiar and age-appropriate contexts. Thus, the test tasks are designed to resemble real-life language use situations that students are likely to encounter in learning English, as well as measure enabling language knowledge and skills that support the development of communication ability. The test scores are determined by the number of questions a student has answered correctly. The number of correct responses on each section is then converted to a scaled score of 101–109 points for Step 1 and 104–115 for Step 2. Responses to the TOEFL Primary Speaking test are scored by human raters at ETS using scoring rubrics with either a 0- to 3-point scale or a 0- to 5-point scale, depending on the task type. The range of speaking scores is 0 to 27. In addition to the numeric scores, score reports contain information in the form of band levels and descriptors to offer teachers information about test-taker performance. For a detailed description of the design of the tests and the interpretation and use of the scores, readers can refer to Cho et al. (2016), ETS (2019), and Papageorgiou and Baron (2017).

Research Questions

Our study addressed the following research questions.

1. Is the difficulty of the test appropriate for students attending the VIPKid MC Levels 3–7?
2. Do students demonstrate a higher level of English proficiency (as measured by the TOEFL Primary tests) relative to their progression through the VIPKid MC Levels 3–7?
3. To what extent do the interactive slides, which contain learning activities and teaching tips, for VIPKid MC Levels 3–7 reflect the content of TOEFL Primary tests?

We collected data on VIPKid students' test performance to answer the first two questions and examined the degree of alignment between the content of the test and the learning activities in the MC curriculum to address the third research question. The data we collected are described next.

Methodology

Collection of Students' Test Score Data

Data collection was conducted in China with VIPKid students taking the TOEFL Primary Listening and Reading tests (i.e., either Step 1 or Step 2) in 2018 and 2019. Due to limited time and resources, a convenience sample strategy (Alreck & Settle, 2004) was used to select the study participants: 5,446 students took the TOEFL Primary Listening test and 5,440 took the TOEFL Primary Reading test. The total number of test takers across the two tests varied because not all students who completed the listening section of the test completed the reading section (four Level 3 students, one Level 4 student, and two Level 5 students). The total number of participants grouped by curriculum level and the TOEFL Primary is summarized in Table 1.

Table 1. Study Participants by Curriculum Level and Test

VIPKid MC level	TOEFL Primary Listening test section			TOEFL Primary Reading test section		
	Step 1	Step 2	Total	Step 1	Step 2	Total
3	253	21	274	251	19	270
4	2,130	367	2,497	2,130	366	2,496
5	608	1,304	1,912	611	1,299	1,910
6	80	581	661	79	583	662
7	8	94	103	8	94	103
Total	3,079	2,367	5,446	3,079	2,361	5,440

Note. MC = Major Course.

Among the participants who took the TOEFL Primary Step 1 Listening test, 1,328 (43.13%) were female, 1,497 (48.62%) were male, and 254 (8.25%) did not report gender. For the TOEFL Primary Step 2 Listening test, 1,016 (42.92%) were female, 1,169 (49.39%) were male, and 182 (7.69%) did not report gender. For the TOEFL Primary Step 1 Reading test, 1,330 were female (43.20%), 1,496 were male (48.59%), and 253 (8.22%) didn't report gender. For TOEFL Primary Step 2 Reading, 1,013 were female (42.91%), 1,166 were male (49.39%), and 182 (7.71%) did not report gender. The mean age of the students by MC level and TOEFL Primary step was between 9.8 and 11.4 (Table 2). Participants took the test in 11 different cities in China. All participants spoke Mandarin as their first language. To address the first two research questions, we calculated descriptive statistics across MC levels and generated histograms and boxplots, which are presented in detail in the Results section.

Table 2. Study Participants by Curriculum Level and Age

VIPKid MC level	TOEFL Primary Listening score						TOEFL Primary Reading score					
	Step 1			Step 2			Step 1			Step 2		
	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*	Min	Max	*M*
3	6	13	10.2	6	12	9.8	6	13	10.2	8	12	9.9
4	6	17	10.7	6	14	10.9	6	17	10.7	6	14	10.9
5	6	15	10.9	6	15	11.2	6	15	10.9	6	15	11.2
6	6	14	11.3	6	15	11.3	6	14	11.3	6	15	11.3
7	8	13	11.4	8	14	11.1	8	13	11.4	8	14	11.1

Note. MC = Major Course.

Data Collection for Content Alignment

For the purpose of analyzing content overlap between the interactive slides and the test tasks, we selected all "entry point" units of MC Levels 3–7, including Units 1, 4, 7, and 10 of Levels 3–6 and Units 1 and 4 for Level 7. Level 1 and Level 2 content was excluded from the study because learners of these two levels are not the target population for the TOEFL Primary due to their young age. We selected the entry point units because they are evenly distributed across the VIPKid MC and they are starting points where new students may begin taking classes based on an in-house placement test. Recall that each unit in Levels 3–6 consists of 12 lessons, which are grouped into two 6-lesson, thematically linked learning cycles: one cycle for Lessons 1–6 and one cycle for Lessons 7–12. For our analysis we selected every other learning cycle, either Lessons 1–6 or Lessons 7–12, of each entry unit. This selection resulted in a total of 120 lessons of the MC curriculum that represented 17% of the lessons from Levels 3–7.

The Hsieh (2015) was used to examine the content alignment of the VIPKid MC interactive slides with the TOEFL Primary tests. For the purposes of this exploratory investigation, we selected three relevant categories for coding (see Table 3). These categories included language skills (listening, reading, speaking), communication goals measured by the test tasks, and the task type (see Cho et al., 2016, for more details on these categories).

Table 3. Coding Categories From the TOEFL Primary Alignment Manual Used in This Study

Language skill	Communication goal	Task type
Listening	Understand simple teacher talks on academic topics	L_Academic Monologue
Listening	Understand dialogues or conversations	L_Dialogue
Listening	Understand spoken directions and procedures	L_Follow Instructions
Listening	Understand simple descriptions of familiar people and objects	L_Listen and Match
Listening	Understand spoken stories	L_Narrative
Listening	Understand dialogues or conversations	L_Question-Response
Listening	Understand short informational texts related to daily life (e.g., phone messages, announcements)	L_Social-Navigational Monologue
Reading	Understand short personal correspondence (e.g., letters)	R_Correspondence
Reading	Understand written expository or informational texts	R_Expository
Reading	Understand written directions and procedures	R_Instructional
Reading	Identify people, objects, and actions	R_Match picture to sentence
Reading	Identify people, objects, and actions	R_Match picture to word
Reading	Understand simple, written narratives (e.g., stories)	R_Narrative
Reading	Understand written expository texts	R_Sentence clue
Reading	Understand commonly occurring nonlinear written texts (e.g., signs, schedules)	R_Telegraphic
Speaking	Ask and answer questions	S_Ask questions
Speaking	Describe people, objects, animals, places, and activities	S_Description
Speaking	Express basic emotions and feelings	S_Expression
Speaking	Make simple requests	S_Requests
Speaking	Explain and sequence simple events	S_Sequence of events
Speaking	Give short commands and directions	S_Short commands and directions

Three VIPKid staff members conducted the coding of the interactive slides. Coder A had previously worked at ETS as a test development specialist for the TOEFL Primary tests for 9 years and was thus knowledgeable of the test content and construct operationalization. Coder A first became familiar with the manual and the MC curriculum and prepared the materials for training and coding purposes. Then, Coder A conducted a training session for Coders B and C, which described the purpose of the study, provided an overview of the test content, and

explained the alignment criteria to be used in the study. The coders first completed coding Level 4 Lesson 1 and Level 4 Lesson 2 together and then coded Level 4 Lesson 3 individually. After a break, coders reconvened and compared their coding for Level 4 Lesson 3 to help ensure consistency and shared understanding of the codes. After this norming session, Level 3 Lesson 1 was coded independently. Coder agreement for Level 4 Lesson 3 was at 62% after completing the norming session, and consensus coding for Level 3 Lesson 1 was at 84%.

After completing the training sessions, all coders coded the remaining 116 lessons. Every coder was assigned to a subset of the interactive slides of the lessons selected for coding. The process for coding the interactive slides of each lesson was as follows:

- Per lesson, the coder reviewed the learning activities and lesson tips of every interactive slide. Using a coding spreadsheet, the coder categorized every interactive slide within a lesson.
- Each interactive slide was coded with the corresponding language skills.
- The coder determined if the learning activity in the slide reflected a TOEFL Primary communication goal and coded the interactive slide appropriately.
- The coder determined if the learning activity in the slide reflected a TOEFL Primary task type and coded the interactive slide appropriately.

To help ensure consistency in the coding process, 22.5% of all selected lessons were double- or triple-coded. Table 4 summarizes all lessons by individual coders, pairs of coders, or the entire group and agreement in terms of percentage of codes applied. Once the coding was completed, discrepancies were resolved in subsequent consensus coding sessions. Coder A and Coder B reviewed the entire coding data set again to make sure that all interactive slides were coded appropriately as intended. Although no universally accepted benchmark regarding the percentage of exact agreement is available (Saldana, 2009), 80% has been proposed as a rule of thumb (McHugh, 2012). Given that our coding process included multiple rounds of coding with consensus sessions to resolve discrepancies, we believe that coding agreement in our study is acceptable.

Table 4. Coders Assigned to Interactive Slides of Unique Lessons

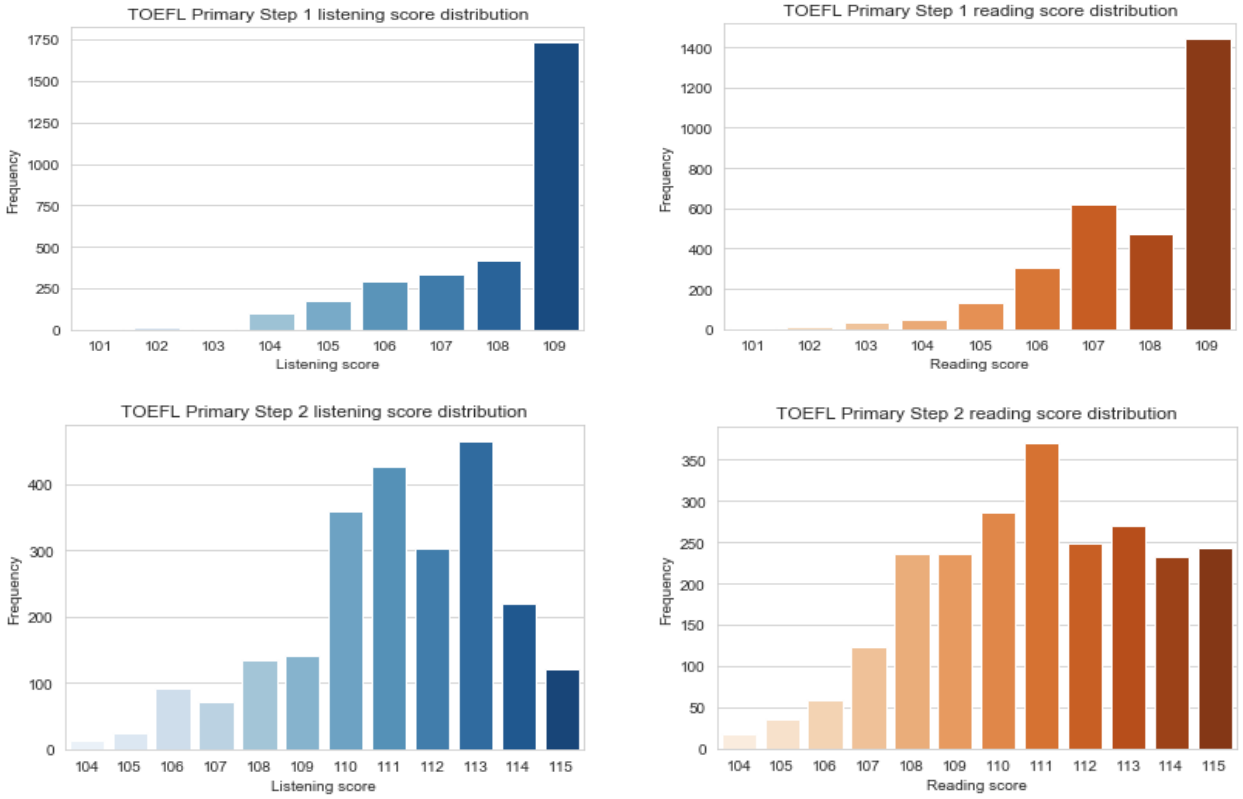
Coder	Count of lessons	Agreement
A	36	
B	37	
C	20	
A + B	8	82%
A + C	7	79%
B + C	7	73%
A + B + C	5	71%
Total unique lessons	120	

Results

Students' Performance on the TOEFL Primary Tests

The distribution of the test scores of all students is presented in Figure 3. For TOEFL Primary Step 1 (Figure 3, top histograms), the distribution was negatively skewed (skewness was -1.45 and -1.30 for listening and reading, respectively), with more scores clustering on the right side of the histograms. Moreover, a clear ceiling effect was observed, as the maximum score on the reported scale, 109, was also the most frequent score.

Figure 3. Distribution of Listening and Reading Scores



Compared to TOEFL Primary Step 1, the distribution of scores for TOEFL Primary Step 2 (Figure 3, bottom histograms) was closer to a normal one (skewness was -0.61 and -0.25 for listening and reading, respectively), without the strong ceiling effect observed with TOEFL Primary Step 1.

Table 5 presents descriptive statistics, specifically mean and standard deviation, by MC level for both sections of TOEFL Primary Steps 1 and 2. Mean score increased as the MC level increased; however, the standard deviation indicated that scores overlapped across MC levels.

Table 5. Descriptive Statistics by Curriculum Level and Test Section

VIPKid MC level	TOEFL Primary Step 1						TOEFL Primary Step 2					
	Listening			Reading			Listening			Reading		
	N	Mean	*SD*	*N*	Mean	*SD*	*N*	Mean	*SD*	*N*	Mean	*SD*
3	253	106.6	2.0	251	106.4	1.9	21	109.3	2.8	19	108.9	3.0
4	2130	107.9	1.5	2130	107.8	1.4	367	109.6	2.6	366	109.4	2.6
5	608	108.4	1.2	611	108.2	1.2	1304	111	2.2	1299	110.7	2.5
6	80	108.8	0.6	79	108.6	0.8	581	112.1	1.9	583	112.1	2.3
7	8	109	0.0	8	108.8	0.7	94	113	1.8	94	113	2.1

Note. MC = Major Course. TOEFL Primary Step 1 scores range from 101 to 109. TOEFL Primary Step 2 scores range from 104 to 115 (ETS, 2019).

Percentile ranks by MC level are presented in Table 6 for TOEFL Primary Step 1 and in Table 7 for TOEFL Primary Step 2 and further confirm earlier observations. For example, the top score of 109 for TOEFL Primary Step 1 was at the 50th percentile of MC Level 4 for the listening section and at the 50th percentile of Level 5 for the reading section. Therefore, TOEFL Primary Step 1 seems to be easy for most students, particularly at MC Level 4 or higher.

Table 6. Percentile Ranks by Curriculum Level for TOEFL Primary Step 1

VIPKid MC level	Listening percentile rank					Reading percentile rank				
	10	20	50	80	90	10	20	50	80	90
3	104	105	107	109	109	104	105	107	108	109
4	106	107	109	109	109	106	107	108	109	109
5	107	108	109	109	109	107	107	109	109	109
6	108	109	109	109	109	107	108	109	109	109
7	109	109	109	109	109	107	109	109	109	109
All	106	107	109	109	109	106	107	108	109	109

Note. MC = Major Course.

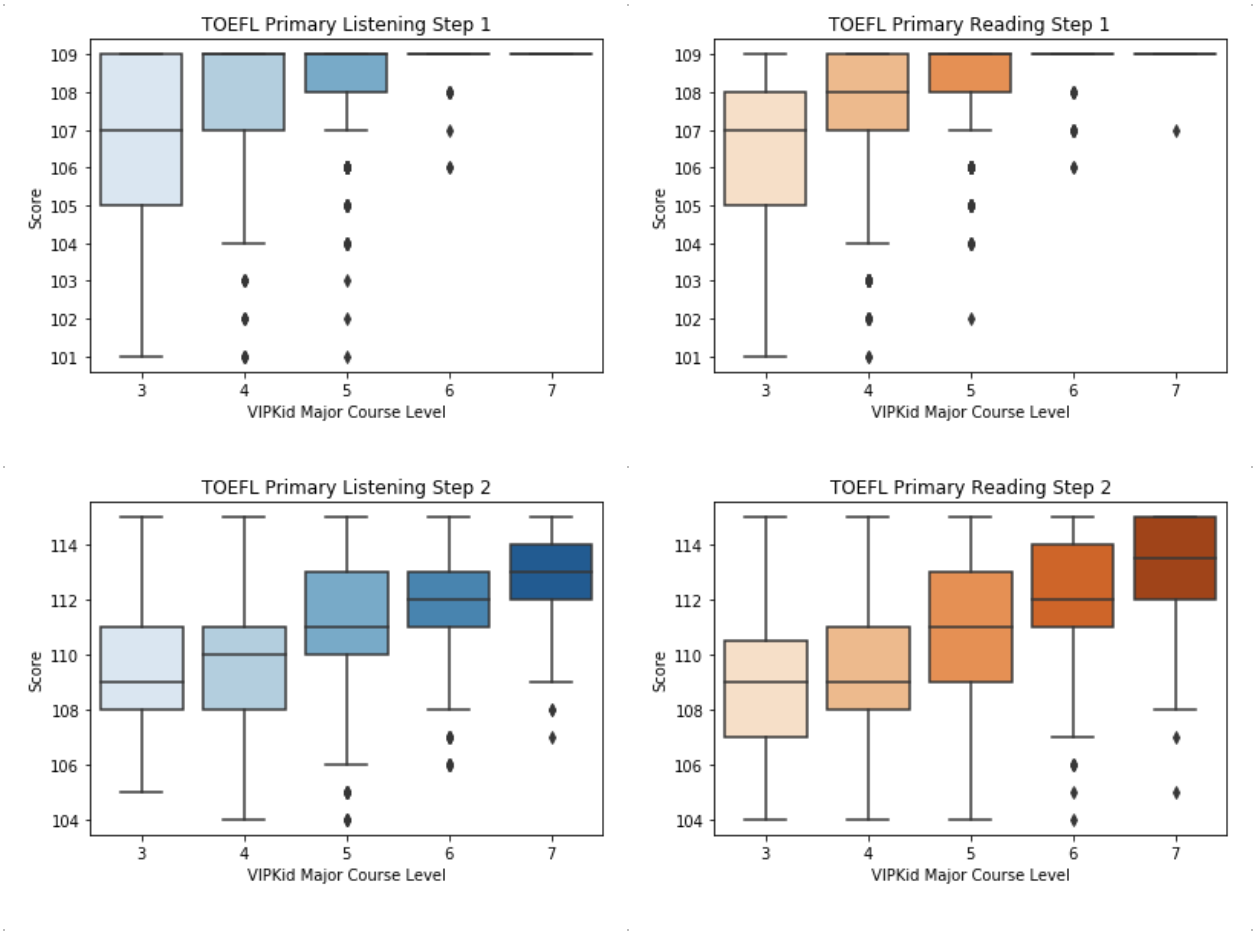
Table 7. Percentile Ranks by Curriculum Level for TOEFL Primary Step 2

VIPKid MC level	Listening percentile rank					Reading percentile rank				
	10	20	50	80	90	10	20	50	80	90
3	106	106	109	113	114	105	106	109	111	115
4	106	107	110	112	113	106	107	109	112	113
5	108	109	111	113	114	107	108	111	113	114
6	110	111	112	114	114	109	110	112	114	115
7	110	112	113	115	115	110	111	114	115	115
All	108	109	111	113	114	108	109	111	114	115

Note. MC = Major Course.

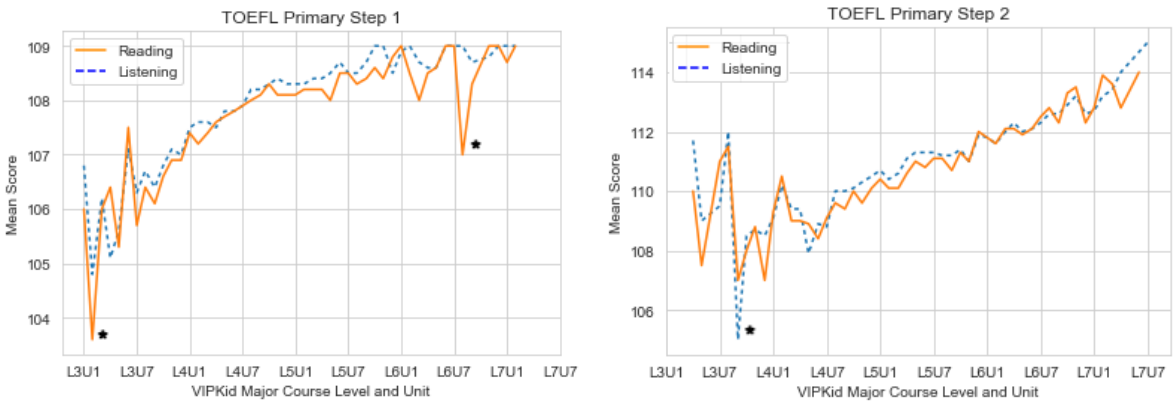
The score patterns detailed in Table 6 and Table 7 are depicted in boxplots in Figure 4. The upper and lower extremes show that the range of test scores was quite large for all MC levels, indicating that some students at low MC levels scored higher than students at high MC levels and vice versa. It should be noted that the outliers in the data had a strong impact on the score ranges; to address this impact, the interquartile range (IQR) is represented by the boxes in Figure 4. IQR is the range for the middle 50% of the scores, as it equals the distance between the first and the third quartiles, which is the distance between the 25th and the 75th percentile; therefore, a key property of the IQR is that it minimizes the distorting effect of extreme scores (Witte & Witte, 2015). With the exception of TOEFL Primary Step 1 test at Level 3, the IQRs were relatively small for all other MC levels, specifically less than half of the score range per TOEFL Primary test. The IQR also trended upward through MC levels. This trend was particularly obvious with TOEFL Primary Step 2 tests, given the strong ceiling effect with TOEFL Primary Step 1 tests. The strong ceiling effect was shown by the median scores, indicated by the continuous horizontal line. The median was the top score of 109 at MC Level 4 for listening and MC Level 5 for reading.

Figure 4. Boxplots of Listening and Reading Scores



In the final analysis of test performance data, we considered the number of units per MC level. As noted earlier, there are a total of 12 units at each MC level (Figure 1), except for Level 7, which has six units. Students typically take one to one-and-half years to complete each level. Figure 5 shows the mean reading and listening scores based on the MC level unit at the time the students took the test. For example, L5U7 on the horizontal axis corresponds to Level 5 Unit 7 (the middle unit of that MC level). Overall, Figure 5 indicates growth of scores from lower to higher MC levels. One apparent exception to this pattern was seen with the TOEFL Primary Step 1 Reading test at MC Level 3 Unit 1 and Level 6 Unit 8. However, there were very few data points at these locations (*N* = 4 and *N* = 1, respectively), creating a disproportionate influence on the graph. Another deviation was seen across Level 3 for TOEFL Primary Step 2 where the total *N* count for the level, regardless of test, was below 25 students (Table 5). These lower *N* counts tended to overly influence the representation of mean scores on the plot.

Figure 5. Mean Scores by Units of Major Course Levels 3–7



Note. L = lesson; U = unit. Asterisks indicate low numbers of test takers attending a lesson unit.

Content Alignment Between the VIPKid Major Course Interactive Slides and the Test Tasks

Interactive slides included in Levels 3–7 of the MC curriculum were coded with regard to the language skills that are covered by the TOEFL Primary tests, that is, listening, reading, and speaking. Table 8 provides an overview of the total number of lesson activities included in each of the five MC levels as well as the number of language-related activities coded. For example, out of 710 activities in Level 3, 571 activities (i.e., 80%) were geared toward teaching and practicing language-related skills. In other words, the MC curriculum has a fairly broad scope insofar as it predominantly features language-related activities while also providing content-related aspects such as math and presentational skills to support language learning.

Table 8. Number of Activities in VIPKid Major Course (MC)

VIPKid MC level	Number of activities	Number of language-related activities (%)
Level 3	710	571 (80.42)
Level 4	663	511 (77.07)
Level 5	587	503 (85.69)
Level 6	578	458 (79.24)
Level 7	740	653 (88.24)
Total	3,278	2,696 (82.25)

As stated earlier, we coded and tallied the interactive slides in the selected units of the MC curriculum relative to the three language skills assessed on the TOEFL Primary tests. The frequencies of the codes are presented in Tables 9 and 10. The category “other” includes

enabling language skills such as phonics, grammar, and vocabulary. The results show that the language skills covered by the TOEFL Primary tests were also covered throughout Levels 3–7 of the MC curriculum. Additionally, findings suggest that the MC curriculum tended to implement both an integrated language skills approach along with a discrete language skills approach. Among the three discrete language skills, reading appeared most frequently toward the higher levels when reading literacy was emphasized in the curriculum.

Table 9. Frequencies of Discrete Language Skills Covered in the Interactive Slides

Language skill	MC Level 3	MC Level 4	MC Level 5	MC Level 6	MC Level 7
Listening only	30	41	12	10	2
Reading only	65	87	122	128	187
Speaking only	48	47	71	52	66
Other	0	19	1	0	7
Total	143	194	206	190	262

Note. MC = Major Course. "Other" in the language skills column refers to enabling skills such as grammar and vocabulary that underlie the three language skills that constituted the focus of the coding exercise.

Table 10. Frequencies of Integrated Language Skills Covered in the Interactive Slides

Language skill	MC Level 3	MC Level 4	MC Level 5	MC Level 6	MC Level 7
Listening + reading	28	49	7	78	7
Listening + speaking	111	73	53	14	23
Reading + speaking	40	32	35	28	39
Listening + other	27	37	12	7	2
Reading + other	175	108	166	129	287
Speaking + other	47	18	24	12	33
Total	428	317	297	268	391

Note. MC = Major Course. "Other" in the language skills column refers to enabling skills such as grammar and vocabulary that underlie the three language skills that constituted the focus of the coding exercise.

The language-related activities in the lessons were also coded in relation to two key design features of the TOEFL Primary tests: communication goals and task types (Cho et al., 2016; ETS, 2019). Table 11 presents the frequencies of the TOEFL Primary communication goals that were present across MC Levels 3–7. It should be noted that not all language-related activities could be further coded to a TOEFL Primary communication goal. For example, a phonics-based activity where a student listens to target phonemes and words and then

practices producing the sounds might be coded to the listening and speaking skills. However, this type of activity did not have a direct corresponding TOEFL Primary communication goal; therefore, it was not accounted for in Table 11.

Table 11. Frequencies of Communication Goals Codes Covered in the Interactive Slides

Language skill	TOEFL Primary communication goal	MC Level 3	MC Level 4	MC Level 5	MC Level 6	MC Level 7
Listening	Understand simple descriptions of familiar people and objects	24	9	0	2	0
Listening	Understand spoken directions and procedures	5	6	0	3	5
Listening	Understand dialogues or conversations	33	47	50	0	15
Listening	Understand spoken stories	0	6	0	1	2
Listening	Understand simple teacher talks on academic topics	2	18	2	2	0
Listening	Total Listening	64	86	52	8	22
Reading	Identify people, objects, and actions	50	15	10	11	11
Reading	Understand simple, written narratives (e.g., stories)	16	12	29	52	68
Reading	Understand written expository or informational texts about familiar people, objects, animals, and places	24	23	41	104	164
Reading	Total Reading	90	50	80	167	243
Speaking	Ask and answer questions	30	19	4	0	23
Speaking	Describe people, objects, animals, places, and activities	75	83	76	52	84
Speaking	Explain and sequence simple events	1	4	9	6	0
Speaking	Express basic emotions and feelings	1	10	18	6	11
Speaking	Total Speaking	107	116	107	64	118
All skills	Grand total	261	252	239	239	383

Note. MC = Major Course.

The frequencies in Table 11 suggest that there was good coverage of all three language skills and communication goals across the MC levels reviewed. It should be noted that some communication goals (e.g., *Understand dialogues or conversations* [Listening], *Understand written expository or informational texts about familiar people, objects, animals, and places* in [Reading], and *Describe people, objects, animals, places, and activities* [Speaking]) have higher

frequencies than others. This coverage may reflect VIPKid's one-to-one style curriculum, which utilizes conversation and discussion and puts an emphasis on oral language use.

Although 12 of the TOEFL Primary communication goals appeared to be well represented across Levels 3–7 in the MC curriculum, six communication goals had no frequency counts and were not included in Table 10. Three of these communication goals—*Understand short personal correspondence* (Reading), *Give short commands and directions* (Speaking), and *Make simple requests* (Speaking)—were represented in other units or levels of the VIPKid MC curriculum, which were not included in this content review. For example, both Levels 1 and 2 of the MC curriculum, which were not included in our review, contain multiple lessons with activities addressing giving short commands and directions as well as making simple requests. Level 3, which was reviewed, contains two units with interactive slides targeting short personal correspondence, but neither of these units were selected for review. Two communication goals—*Understand short informational texts related to daily life* (Listening), and *Understand written directions and procedures* (Reading)—were not included as activities in the interactive slides we analyzed; however, it is reasonable to expect that such communication goals will appear in student–teacher interactions on the platform or in pre- and postclass materials. For example, students and teachers routinely begin each class by discussing their weekend activities or conversing about other personal information. However, these types of common rapport-building interactions are not specifically detailed within the interactive slides of a lesson and, therefore, were not counted toward the communication goal *Understand short informational texts related to daily lif*e (see the Discussion and Conclusions section). Within pre- and postclass materials, which also were not analyzed for this study, students read and follow directions and procedures presented on-screen. Lastly, the reading communication goal *Understand commonly occurring nonlinear written texts* was not represented in the MC curriculum. The designers of the curriculum could consider adding more tasks and activities in the future to cover this communication goal, broaden the scope of text types included in the curriculum, and thus provide students with exposure to this type of reading passages.

Table 12 presents the frequencies of the task type codes by each MC level. As with the communication goals, it was not possible to code some language-related activities in relation to

a TOEFL Primary task type. Returning to the previous example of a phonics activity, there was no corresponding TOEFL Primary task for that type of practice activity.

Table 12. Frequencies of Task Types Covered in the Interactive Slides

Language skill	TOEFL Primary task type	MC Level 3	MC Level 4	MC Level 5	MC Level 6	MC Level 7
Listening	Listen and match (Step 1)	24	9	0	2	0
Listening	Question-response (Step 1)	23	25	2	0	0
Listening	Dialogue (Both Step 1 and Step 2)	10	22	48	0	15
Listening	Follow instructions (Both Step 1 and Step 2)	5	6	0	3	5
Listening	Narrative (Step 2)	0	6	0	1	2
Listening	Academic monologue (Step 2)	2	18	2	2	0
Listening	Total Listening	64	86	52	8	22
Reading	Match picture to word (Step 1)	42	8	9	8	8
Reading	Match picture to sentence (Step 1)	8	7	1	3	3
Reading	Sentence clue (Both Step 1 and Step 2)	24	18	27	40	60
Reading	Narrative (Step 2)	16	12	28	52	68
Reading	Short expository (Step 2)	0	5	15	64	104
Reading	Total Reading	90	50	80	167	243
Speaking	Ask questions	30	19	4	0	23
Speaking	Description	74	83	79	52	84
Speaking	Expression	2	10	15	6	11
Speaking	Sequence of events	1	4	9	6	0
Speaking	Total Speaking	107	116	107	64	118
All skills	Grand total	261	252	239	239	383

Note. MC = Major Course.

The frequencies in Table 12 suggest relatively good coverage of task types across all three language skills covered by the TOEFL Primary tests and throughout the MC levels. Some task types (e.g., *Dialogue* for Levels 4–7 [Listening], *Sentence clue* for Levels 3 and 4, and *Narrative* for Levels 5–7 [Reading], and *Description* [Speaking]) had higher frequencies than others in each of the skills coded. This may further reflect VIPKid's curriculum style with emphasis on oral language use and conversation of shared content in one-to-one online settings, starting with building language skills at earlier levels, and longer texts, such as narrative stories, as the curriculum progresses.

Similar to Table 11, Table 12 only includes task types with code frequencies in our data. Task types without any frequencies were *Social-navigational monologue* (Listening), *Correspondence* (Reading), *Telegraphic* (Reading), *Instructional* (Reading), *Requests* (Speaking), and *Short commands and directions* (Speaking). As discussed previously, some of the task types (*Correspondence, Instructional, Requests,* and *Short commands and directions)* were included in the MC curriculum but were not represented in the 17% of reviewed content. For example, requests as well as short commands and directions were a constitutive component of the oral interaction in almost all one-on-one instruction sessions. Similarly, the many interactions during the teaching sessions closely resembled *Social-navigational monologue* tasks but were not featured as stand-alone activities or tasks in the MC curriculum. Finally, though some nonlinear texts did appear in the curriculum, they were used as graphic organizers and not fully completed materials, which did not directly correspond to *Telegraphic* tasks.

Discussion and Conclusions

The purpose of our study is to explore the appropriateness of using the TOEFL Primary tests to evaluate the English language abilities of EFL students attending the VIPKid MC Levels 3 to 7. To address our research questions, we collected data on students' test performance and we examined the degree of alignment between the content of the test and the MC learning activities to address two research questions.

Regarding the first question (*Is the difficulty of the test appropriate for students attending VIPKid MC Levels 3–7?*), analysis of test score data showed that the TOEFL Primary tests are in general at the appropriate level of difficulty for the students. A strong ceiling effect was observed with the TOEFL Primary Step 1 tests starting at MC Level 4, in particular with the listening test. Regarding the second question (*Do students demonstrate a higher level of English proficiency [as measured by the TOEFL Primary tests] relative to their progression through the VIPKid MC Levels 3–7?*), score data analysis revealed an overall tendency for scores to grow across the curriculum levels.

The score data findings have implications for the use of the TOEFL Primary tests students who learn English with the VIPKid curriculum. In terms of difficulty, the TOEFL Primary Step 1 tests might be more appropriate for MC Level 3 students and the TOEFL Primary Step 2

tests more appropriate for students at MC Levels 5–7. For MC Level 4 students, the choice of tests might be based on instructional goals. For example, the TOEFL Primary Step 1 test might be preferred over the TOEFL Primary Step 2 test when the goal is to offer students a positive experience through achieving a score close to the top of the score scale. When the goal is to help students transition to more challenging tests, then the TOEFL Primary Step 2 tests might be the preferred option. The choice of the tests can also be informed by administering a short screener test designed for TOEFL Primary students (Schmidgall et al., 2018).

Regarding the third research question (*To what extent do the interactive slides, which contain learning activities along with teaching tips, for VIPKid MC Levels 3–7 reflect the content of TOEFL Primary tests?*), the findings from the systematic coding indicated an overall good match between the interactive slides of entry point units of MC Levels 3–7 and what is being assessed on the TOEFL Primary tests. For example, as shown in Table 9, all language-related skills measured by TOEFL Primary tests are also included across MC Levels 3–7 in both discrete and integrated activity formats. Additionally, the frequencies of language skills included in the MC interactive slides highlight an initial focus of oral language skills such as listening and speaking as well as a gradual increase in target language literacy skills such as reading—a structure that reflects the conceptual underpinning and design considerations behind the TOEFL Primary tests (Cho et al., 2016).

Moreover, the gradual increase in the number of language activities related to reading is a trend that seems to parallel the distribution of reading communication goals on the TOEFL Primary Step 1 and Step 2 tests. For example, as Table 10 and Table 11 show, we found communication goals and task types that tend to be associated with TOEFL Primary Step 1 (e.g., *Identify people, objects and actions* and *Match picture to word)* mainly in the lower MC levels. By contrast, goals associated with more complex reading passages, featured particularly in TOEFL Primary Step 2 (e.g., *Understand simple, written narratives [e.g., stories]* and *Understand written expository or informational texts about familiar people, objects, animals, and places*), were found more often at the higher levels of the MC curriculum. Hence, although we did not directly code the interactive slides in terms of complexity, there is preliminary evidence that core features of the content taught in the MC levels align with the design of the TOEFL Primary

tests as there is a gradual increase in the complexity of the communication goals found across TOEFL Primary Step 1 and Step 2.

Finally, there appears to be a good coverage of task types. As shown in Table 11, most task types on the TOEFL Primary tests featured quite prominently among the interactive slides of MC Levels 3–7. As we discussed earlier, a good representation and overlap of task types between learning activities and test tasks is particularly crucial for YLs. Such representation and overlap facilitate the familiarization of YLs with tasks included in an assessment and help avoid confusion or anxiety in children (Hill & Wigfield, 1984; Lee & Winke, 2018). In other words, the fairly strong representation of task types across the curriculum may even serve a positive effect insofar as they familiarize the YLs with the test tasks and help prepare them for the test-taking experience.

To summarize, the study found a satisfactory alignment between the interactive slides used in the entry point units of VIPKid MC and the tasks in the TOEFL Primary tests. To a large extent, the language skills, communication goals, and task types featured on the tests match the ones underlying and deployed in the English instruction provided by VIPKid through the interactive slides. However, it needs to be noted that some TOEFL Primary communication goals and task types were not found among the interactive slides in Levels 3–7. As mentioned above, aspects of VIPKid's content were not included in the coding, including teacher–student interactions and pre- and postclass materials. Thus, in order to obtain a more complete picture of the alignment between the TOEFL Primary tests and the VIPKid MC curriculum than offered in this study, future studies may need to systematically account for student–teacher interactions during the actual online classes. Taking into account interactions and activities conducted during the learning sessions might reveal more comprehensive alignment between the curriculum, the goals, and the task types of the TOEFL Primary tests and the VIPKid curriculum than was found in our study.

Naturally, our study comes with some limitations. First, due to time and human resource constraints, we limited our content analysis to the interactive slides of entry point units, due to their consistent distribution across the VIPKid MC. Second, we were not able to conduct teacher focus groups or interviews as was the case with similar alignment studies in the context

of TOEFL YSS (Hsieh, 2015; Timpe-Laughlin, 2018). In addition to teacher perspectives, future alignment research may also want to collect the perspectives of YLs who took the test as part of their learning an additional language. As shown by Butler et al. (2020), YLs already have a considerable amount of assessment literacy. Moreover, they are the major stakeholders in the assessment process. Hence, including their perspective may not only offer a valuable, additional piece of evidence in an alignment study, but also provide students with agency in their own journey of learning an additional language (Hsieh & Gu, 2020). Thus, future alignment studies should consider including both teacher and student perspectives as these stakeholder groups can offer additional insights into the appropriateness of using TOEFL Primary tests to evaluate the language proficiency of students attending MC Levels 3–7. Finally, we did not embark on a systematic investigation of the student–teacher interactions during the online classes. However, as learning increasingly shifts online, future alignment studies might also consider the unique characteristics of providing EFL instruction to YLs in the context of an online classroom.

Despite the above limitations, we believe that our study makes a modest contribution to the line of alignment research conducted between the TOEFL YSS and English curricula in different geographical contexts around the world. Whereas previous studies have investigated EFL curricula in Africa (Hsieh, 2015; Hsieh et al., 2018) and Europe (Timpe-Laughlin, 2018), this study shows alignment with a different type of English curriculum for YLs of English in Asia. Consequently, our study offers additional empirical evidence supporting the usefulness of the TOEFL Primary tests in specific curricular settings, given that the TOEFL Primary tests were not designed based on a specific curriculum. Beyond the specific context we investigated, we also believe that our research contributes to the field of EFL YL research by demonstrating a systematic approach to exploring the appropriateness of using an external assessment within a specific educational context. As we noted in our review of the relevant literature, a decision to use an external assessment should be supported by evidence of its appropriateness for a given students population, in particular, young EFL learners, because of their rapid cognitive, social, and affective development.

References

Alreck, P. L., & Settle, R. B. (2004). *The survey research handbook.* McGraw-Hill/Irwin.

Barrouillet, P. (2015). Theories of cognitive development: From Piaget to today. *Developmental Review*, *38*, 1–12. https://doi.org/10.1016/j.dr.2015.07.004

Bishop, J. L., & Verleger, M. A. (2013, June 23–26). *The flipped classroom: A survey of the research* [Paper presentation]. 2013 ASEE Annual Conference & Exposition, Atlanta, GA, United States. https://peer.asee.org/22585

Butler, Y. G., Peng, X., & Lee, J. (2020). *"I want humanized assessment!": Young learners' voices for a learner-centered approach to language assessment literacy* [Manuscript submitted for publication].

Cameron, L. (2001). *Teaching languages to young learners*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733109

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). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-16-02.pdf

Council of Europe. (2001). *The Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/1680459f97

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for languages: Learning, teaching, assessment. A manual.* Cambridge University Press. https://rm.coe.int/1680667a2d

Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences, 19*(2) 209–218. https://doi.org/10.1016/j.lindif.2009.01.002

Deville, C., & Chalhoub-Deville, M. (2011). Accountability-assessment under No Child Left Behind: Agenda, practice, and future. *Language Testing, 28*(3), 307–321. https://doi.org/10.1177/0265532211400876

Educational Testing Service. (2019). *TOEFL research insight series: Vol. 8. TOEFL Primary® framework and test development.* Educational Testing Service. https://www.ets.org/toefl/research/insight-series

Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. Cito; Council of Europe; European Association for Language Testing and Assessment.

Fisher, D., & Frey, N. (2013). *Better learning through structured teaching: A framework for the gradual release of responsibility* (2nd ed.). ASCD.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Lawrence Erlbaum Associates.

Harsch, C., & Hartig, J. (2015). What are we aligning tests to when we report test alignment to the CEFR? *Language Assessment Quarterly*, *12*(4), 333–362. https://doi.org/10.1080/15434303.2015.1092545

Herman, J. L., & Webb, N. M. (2007). Alignment methodologies. *Applied Measurement in Education, 20*(1), 1–5. https://doi.org/10.1080/08957340709336727

Hill, K. T., & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. *The Elementary School Journal, 85*(1), 105–126. https://doi.org/10.1086/461395

Hsieh, C.-N. (2015). *Examining the use of TOEFL Primary for young English learners in the context of English-medium instruction* [Unpublished manuscript].

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

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

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. T*heory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing*, *35*(2), 239–269. https://doi.org/10.1177/0265542217704 9

Martyniuk, W. (Ed.). (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge University Press.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Nikolov, M., & Timpe-Laughlin, V. (2020). Assessing young learners of foreign languages. *Language Teaching.* Advance online publication. https://doi.org/10.1017/S0261444820000294

No Child Left Behind Act. 20 U.S.C. § 6301. (2001).

Papageorgiou, S. (2016). Aligning language assessments to standards and frameworks. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 327–340). Mouton de Gruyter.

Papageorgiou, S., & Baron, P. (2017). Using the Common European Framework of Reference to facilitate score interpretation for young learners' English language proficiency assessments. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 136–152). Routledge. https://doi.org/10.4324/9781315674391-8

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-15-06.pdf

Papageorgiou, S., Wu, S., Hsieh, C.-N., Tannenbaum, R. J., & Cheng, M. (2019). *Mapping the* TOEFL iBT® *test scores to China's Standards of English Language Ability: Implications for score interpretation and use* (TOEFL Research Report No. 89). Educational Testing Service. https://doi.org/10.1002/ets2.12281

Saldana, J. (2009). *The coding manual for qualitative researchers* (2nd ed.). Sage.

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools*, *45*(2), 158–176. https://doi.org/10.1002/pits.20282

Samana, W. (2013). Teacher's and students' scaffolding in an EFL classroom. *Academic Journal of Interdisciplinary Studies*, *2*(8), 338–343. https://doi.org/10.5901/ajis.2013.v2n8p338

Schmidgall, J. E., Getman, E. P., & Zu, J. (2018). Screener tests need validation too: Weighing an argument for test use against practical concerns. *Language Testing*, *35*(4), 583–607. https://doi.org/10.1177/0265532217718600

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting approaches to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, *11*(3), 233–249. https://doi.org/10.1080/15434303.2013.869815

Timpe-Laughlin, V. (2018). *A good fit? Examining the alignment between the TOEFL Junior® Standard test and the English as a foreign language curriculum in Berlin, Germany.* (Research Memorandum No. RM-18-11). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-18-11.pdf

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7–25. https://doi.org/10.1080/08957340709336728

Witte, R. S., & Witte, J. S. (2015). *Statistics* (10th ed.). Wiley.

Notes

[1] The interactive slide is provided to illustrate the design of the platform. To protect proprietary content, this interactive slide was not part of the data analyzed in our study.