Examining Content Representativeness of a Young Learner Language Assessment: EFL Teachers' Perspectives

Ching-Ni Hsieh

Abstract This study aims to provide content validity evidence for the new young language learner assessment—TOEFL Primary—a test designed for young learners ages 8 and above who are learning English in English as a Foreign Language (EFL) contexts. The test focuses on core communication goals and enabling language knowledge and skills represented in various EFL curricula. A panel of 17 experienced EFL teachers, representing 15 countries, participated in the study. The teachers evaluated the relevance and importance of the knowledge, skills, and abilities (KSAs) assessed by the reading and listening items of TOEFL Primary. Content Validity Indices (CVIs) (Popham, Appl Meas Educ 5(4):285–301, 1992) was used to determine the degree of match between the test contents and the target constructs and the importance of the KSAs assessed for successful classroom performance. Results showed that the majority of the items had an average CVI above the cut-off value of .80, indicating that the items measured what they were intended to measure and that the KSAs assessed were important for effective classroom performance, supporting the claim about using the test scores to support language teaching and learning.

Keywords Content validity • TOEFL Primary • Young learners • Language assessments • Teacher judgments • Language teaching

1 Introduction

Measuring and reporting content validity of newly developed tests is important because this type of validity evidence provides test users essential information regarding the extent to which test contents reflect the target constructs being

C.-N. Hsieh (✉)
Center for English Language Learning and Assessment, Research and Development,
Educational Testing Service, Princeton, NJ, USA
e-mail: chsieh@ets.org

M. Nikolov (ed.), *Assessing Young Learners of English: Global and Local Perspectives*, Educational Linguistics 25, DOI 10.1007/978-3-319-22422-0_5

measured and the validity of the inferences drawn from the test scores (D'Agostino, Karpinski, & Welsh, 2011; Haynes, Richard, & Kubany, 1995; So, 2014; Yalow & Popham, 1983). The study reported here examines the degree of content representativeness within the context of a new young learner language assessment, TOEFL Primary, with the goal of providing an important piece of content validity evidence for the test.

As the number of young English language learners worldwide continues to grow, so too does the need for language assessments designed to target this population (McKay, 2006; Nikolov, 2016, in this volume). While several language assessments have been developed to serve the needs of these learners (e.g., Cambridge English: Young Learners English Tests; TOEFL Primary; TOEFL Junior), theoretical and empirical knowledge about the assessment of young language learners remains underdeveloped. For instance, relatively little is known about the target language use (TLU) domains for English communication among young learners. What is clear, however, is that language tasks designed for young learners need to take into consideration factors such as learners' shorter attention span (Robert, Borella, Fagot, Lecerf, & De Ribaupierre, 2009), memory capacity (Cho & So, 2014), longer processing time (Berk, 2012), developing literacy, and limited exposure to and experience of the world—factors that are distinct from those relevant to the assessments of adult learners of English as a Second (or Foreign) Language (ESL/EFL). Given these differences, it is critical for language test developers and researchers to better comprehend how the test contents of young learner assessments reflect and meet the communication needs of young learners and how individual characteristics of students should influence test design.

TOEFL Primary is a new young learner language assessment developed by Educational Testing Service (ETS). The test is designed for young learners ages eight and above who are learning English in EFL contexts. The test measures three English language skills: listening, reading, and speaking. Listening and reading are offered in two steps, i.e. Step 1 (low level) and Step 2 (high level), to reflect the wide range of language proficiency exhibited among the target population. The speaking test is designed for language learners at many different proficiency levels of English, from beginners to more proficient speakers, and thus is not separated into different steps. The test items of TOEFL Primary cover a set of communication goals, a range of difficulty, and various item types. The test is intended to support language teaching and learning by providing meaningful information for the test takers' current English ability. EFL teachers can use the test to guide their teaching goals, monitor student progress, and identify students' strengths and weaknesses in different areas of language use. The test scores can also be used for placement purposes if the test content corresponds to or is relevant to the content of the EFL curriculum that the students are exposed to. However, the test is not intended to support high-stakes decisions such as to inform admission decisions or to evaluate teachers' performances.

2 Literature Review

The link between test content and EFL curricula is an important facet in establishing content validity for tests that are developed to provide instructional support. Two studies that examined the relationships between test contents and course contents (Fleurquin, 2003; Wu & Lo, 2011) have specific implications for the current study. Fleurquin reported the process of developing and validating *Alianza Certificate of Elementary Competence in English* (ACECE), a standardized test of American English that measures young learners' English communication skills within the context of elementary schools in Uruguay. To examine content validity of the ACECE, the research team enlisted experienced EFL teachers to compare the grammar structures and vocabulary categories assessed in the test with the contents of three textbooks used with the target population in local schools. The comparison showed that the majority of the grammar structures and vocabulary assessed in the test matched those presented in the textbooks that the students had used during their school years, providing evidence to support the alignment between the content of the ACECE and the three textbooks. Specific comments regarding the test items and stimulus materials provided by the EFL teachers were also used to inform test revisions.

Wu and Lo (2011) investigated the relationship between a standardized English language proficiency test for young children, the Cambridge English: Young Learners English (YLE) Tests, and the EFL teaching practices at the elementary level in Taiwan. The study aimed to inform local teachers regarding whether the YLE tests were suitable for young learners in Taiwan. The researchers compared the Grades 1–9 Curriculum Guidelines published by the Ministry of Education in Taiwan and a popular series of English textbooks published by a local publisher with the content of the YLE. The comparison was conducted in six aspects: topics, grammar and structures, communication functions, competence indicators, vocabulary, and tasks. Results showed a moderate to high degree of alignment between the YLE and the local teaching practices with regard to the six aspects of the comparison and highlighted a gap between the two in terms of cultural differences between Taiwan and the UK as manifested in the wordlists introduced. Taken together, the use of expert teacher judgments in Fleurquin (2003) and Wu and Lo (2011) has proven useful in helping researchers and test developers determine content alignment between young learner language assessments and EFL curricula in different EFL contexts and identify aspects of misalignment to inform test revisions.

It needs to be noted that in content validation studies that use expert judgments, a criterion (i.e., cut-off point) is required to ensure the quality of the judgments. While both Fleurquin (2003) and Wu and Lo (2011) used expert teachers to evaluate the alignment between test content and local teaching practices, neither study employed a definite cut-off value, leaving open a determination of the test's content representativeness. Since one major purpose of content validation studies is to ensure that the test contents reflect what they are intended to measure, a criterion for making that decision is critical to represent the quality of the test content. The more stringent the criterion is, the more confidence that can be placed in positive appraisals of the test content (Popham, 1992).

In this study, I examined the content representativeness of TOEFL Primary using a traditional content validity approach based on the computation of a Content Validity Index (CVI) (Davis, 1992; Lynn, 1986) with a predetermined criterion. The CVI approach entails a panel of expert judges evaluating whether the relevance of each test item on an assessment instrument is relevant to the target construct being measured. The percentage of items rated as relevant by each judge and the average of the percentages across the judges are reported as an indication of the degree of "content validity", or more appropriately, content representativeness in this case. The use of CVIs to determine content representativeness is widely cited in test development literature for teacher licensure tests (Crocker, Miller, & Franks, 1989; Popham, 1992), nursing research (Davis, 1992; Polit & Beck, 2006) and social work research (Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003), but to the best of my knowledge, they have not been widely used for tests of second language proficiency.

3 Content Validation of TOEFL Primary

During the initial stage of test development of TOEFL Primary, the researchers and test developers at ETS had set out to conduct a two-stage process for establishing the content validity of the test (Lynn, 1986; Sireci, 1998, 2007). The first stage, or 'Developmental Stage,' involves the identification of the domain of content through a comprehensive review of relevant literature and domain analysis of language use in EFL classrooms—the TLU domain. The domain descriptions were enhanced by the development team's review of EFL curricula and textbooks used in nine countries, including Brazil, Chile, China, Egypt, Japan, Korea, the Philippines, Qatar, and Singapore (Turkan & Adler, 2011). Results of the domain analysis helped define the construct of English communication for young learners. A set of communication goals that are unique to young learners' communicative needs and the language knowledge and skills required to fulfill these communication goals are incorporated in the construct definitions. The communication goals targeted also helped test developers identify specific text types that young learners encounter in their EFL reading and listening materials and the various types of speaking activities that young learners engage in the EFL classrooms. A variety of test tasks associated with specific communication goals are developed for the test.

The second stage of content validation, the 'Judgment/Quantification' stage of content validation (Lynn, 1986), for TOEFL Primary is twofold, involving a teacher survey on the pilot-test items and a panel judgment of the operational test items—i.e. the current study. During pilot testing of TOEFL Primary, a teacher survey study was conducted at local testing sites where TOEFL Primary was piloted. The survey aimed to gather EFL teachers' feedback on the importance and relevance of the set of communication goals identified for construct definitions and the appropriateness and effectiveness of the task types proposed for young EFL learners. Results of the teacher survey, which contained the evaluations of test contents by 29 EFL teachers from Costa Rica, Egypt, Japan, Peru, and Vietnam, showed that the communication goals substantially reflected the communicative needs of young learners. The survey also revealed varying views regarding the effectiveness of the task types, which subsequently informed the subsequent refinement of the tasks (Hsieh, 2013).

The current study focused on the panel judgment of the TOEFL Primary operational listening and reading items in terms of their content relevance and the importance of the language knowledge, skills, and abilities (KSAs) assessed in these items for successful classroom performance. The study was informed by the body of literature that uses CVIs to determine the degree of content representativeness for newly developed language assessments. Predefined cut-off values suggested by the collective body of literature (e.g., Davis, 1992; Lynn, 1986) were adopted for determining whether test items were congruent with the constructs being measured and whether the KSAs assessed reflected those introduced in a number of EFL contexts. The use of CVIs to assess the degree of agreement among the EFL teachers has the benefit of allowing better comparability between the judgments gathered by different content validity studies.

The study aimed to address the following research questions:

1. To what extent do TOEFL Primary listening and reading test items reflect the target constructs as judged by EFL teachers?
2. What are EFL teachers' perceptions of the importance of the KSAs assessed by TOEFL Primary in their specific teaching contexts?

4 Method

4.1 Participants

A panel of 17 EFL teachers served as the expert judges in this study. The panel of judges was formed, to the extent possible with a relatively small sample, to have representation by gender, professional background, and geographic location. Participants were selected from a large pool of EFL teachers based on their expertise in young learner EFL curricula and professional experience. All teachers had experience teaching young learners similar to the target population for TOEFL Primary, i.e. ages eight and above. Fifteen countries (Brazil, China, France, Greece, Japan, Jordan, Kazakhstan, Mexico, Peru, Russia, Slovakia, South Korea, Spain, Sweden, and Vietnam) were represented. The teachers were between the ages of 25 and 52 (*Mean* = 38.9, *SD* = 7.3). Their years of teaching EFL ranged from 3 to 29 years (*Mean* = 14.9 years, *SD* = 7.0). Table 1 shows the demographic information of the teachers.

Table 1 Demographic information of the participating teachers

Educational background	N	%
College	5	29 %
Some postgraduate education	2	12 %
Master	8	47 %
PhD	2	12 %
Year of teaching		
Below 10 years	4	24 %
10–20 years	9	53 %
21–30 years	4	24 %
Age		
20s	2	12 %
30s	5	29 %
40s	8	47 %
50s	2	12 %
Gender		
Male	4	24 %
Female	13	76 %
Geographical region		
Asia & The Middle East	6	35 %
Europe	7	41 %
Latin America	4	24 %

4.2 Rating Materials

The rating materials used in this study consisted of operational listening ($N=57$) and reading ($N=57$) test items of TOEFL Primary. These items were carefully chosen by the test developers at ETS to cover all the targeted communication goals of TOEFL Primary, the full range of difficulty, and all item types (see Table 2). The number of items per item type reflected that of the operational form. The total number of the listening and reading items included in the study was larger than the number in an operational form because these items covered the two difficulty levels of TOEFL Primary. The inclusion of items from both steps was considered important to ensure a comprehensive coverage of the difficulty range of the test. Including more items in the study was also thought to produce more stable judgments overall. The speaking section was not included in the study due to time and resource constraints in data collection.

4.3 Instrument

A content alignment questionnaire for item evaluation was constructed by the researcher through consultation with ETS test developers and research scientists who were experienced with content alignment studies. The instructions to

Table 2 TOEFL Primary listening and reading items for evaluation

Listening item type	Communication goal	Step	N
Listen and match	Understand simple descriptions of familiar people and objects	1	7
Follow instructions	Understand spoken directions and procedures	1, 2	10
Question/response	Understand dialogues or conversations	1	6
Dialogue	Understand dialogues or conversations	1, 2	10
Social-navigational monologue	Understand short informational texts related to daily life	1, 2	10
Narrative set	Understand spoken narratives	2	8
Academic monologue	Understand expository monologues	2	6

Reading item type	Communication goal	Step	N
Match picture to word	Identify people, objects and actions	1	6
Match picture to sentence	Identify people, objects and actions	1	7
Sentence clues	Understand written expository or informational texts	1, 2	12
Telegraphic sets	Understand commonly occurring non-linear written texts (e.g. signs, schedules)	1, 2	8
Correspondence	Understand short personal correspondence	1, 2	6
Instructional texts	Understand written directions and procedures	2	6
Narrative sets	Understand simple, written narratives	2	8
Expository paragraph	Understand written expository or informational texts about familiar people, objects, animals, and places	2	4

participants during the alignment exercise, the questionnaire response formats and scales underwent multiple rounds of trials and revisions prior to data collection. The final survey instrument consisted of two subsections. Section I included seven parts, each corresponding to one listening item type. Section II included eight parts, each corresponding to one reading item type. The KSAs assessed in each item type were provided in the questionnaire to facilitate the evaluation process.

4.4 Procedures

The 17 EFL teachers were invited from their countries to ETS campus in Princeton, New Jersey, to participate in the study. Each teacher was supplied with (a) a background questionnaire that was used to gather the teachers' biographical information, (b) a test booklet that contained the 57 listening and 57 reading test items, (c) a copy of the scripts for the listening items, and (d) the content alignment questionnaire for the evaluation of the test items. Prior to the day of the content alignment exercise, all teachers took the TOEFL Primary test and reviewed documents on the test design framework and scoring guidelines to become familiar with the test constructs, design, and scoring criteria. On the day of data collection, the teachers first completed the background questionnaire and then were instructed to make judgments on two aspects of the content representativeness of each item using the content alignment questionnaire. The two aspects were content relevance of and the importance of the KSAs assessed by the TOEFL Primary test items. In addition to the content alignment exercise, five teachers (from France, Jordan, Mexico, Peru, and Spain) agreed to participate in follow-up interviews that were conducted after the analyses of the rating data. The interviews focused on (1) the teachers' views about specific aspects of the test contents that the teachers considered less important or relevant to their own teaching practices and (2) how the teachers used the different types of texts and item types in their respective EFL classrooms.

4.5 Content Alignment Judgments

The two aspects of content alignment judgments the teachers were asked to perform are described as follows.

(1) *Content relevance*

The first judgment asked the teachers to evaluate the degree to which the content of each item reflected the target construct it is intended to measure. Congruent with Lynn's (1986) item relevance rating rules, judges were asked to provide the relevance ratings on a Likert scale with four possible responses: *no reflection, slight reflection, moderate reflection* and *strong reflection*. Responses of 'moderate reflection' and 'strong reflection' were regarded as indications of teachers' endorsement of the content relevance of the items, whereas responses of 'no reflection' and 'slight reflection' indicated the opposite. The responses were dichotomized in this fashion in order to facilitate summary evaluations.

(2) *The importance of the KSAs assessed*

The second judgment required the teachers to rate the importance of the KSAs required of young EFL learners for successful classroom performance in their own teaching contexts. The importance ratings, also on a 4-point Likert scale (Lynn, 1986), had four different labels: *not important, somewhat important, important* and *very important.* Responses of 'important' and 'very important' indicated teachers' agreement on the importance of the KSAs assessed, whereas responses of 'not important' and 'somewhat important' indicated the opposite. As with the content relevance ratings, the importance ratings were also dichotomized.

4.6 Analysis

To answer the research questions, individual ratings provided by the 17 judges were pooled and the CVIs for each item were calculated for evaluating the degree of content relevance and importance of the KSAs assessed in the TOEFL Primary test items (Davis, 1992; Lynn, 1986; Polit & Beck, 2006). The analyses of the degree of content representativeness of the test items are described below.

(1) *CVIs for content relevance*

For the content relevance ratings, the CVI for each item was calculated by counting the number of judges who rated that item as either 'moderate reflection' or 'strong reflection' and dividing that number by the total number of judges. The CVI calculated for each item provided information about the proportion of judges who considered an item as content relevant. The CVIs for the listening and reading sections were defined as the proportion of items on the section that achieved a rating of 'moderate reflection' or 'strong reflection' across all judges. The CVIs for listening and reading sections were derived, respectively, by averaging the CVIs across the 57 items for each section.

(2) *CVIs for the importance of the KSAs assessed*

For the importance of the KSAs assessed, the CVI for each item was calculated by counting the number of judges who rated the item as either 'important' or 'very important' and dividing that number by the total number of judges. The CVI calculated for each item provided information about the proportion of judges who considered the KSAs assessed by an item as important for successful classroom performance. The CVIs for the listening and reading sections were defined as the proportion of items on the section that achieved a rating of 'important' or 'very important' across all judges. The CVIs for listening and reading sections were derived, respectively, by averaging the CVIs across the 57 items for each section.

To determine the degree to which TOEFL Primary test items reflect the target constructs and assess the important KSAs required of young learners, a CVI of .80 was used as the acceptable criterion, following Davis (1992). This criterion is widely used in the literature for determining content representativeness of new assessments (e.g., Rubio et al., 2003). This cut-off value indicates that, when a total of 17 judges are considered, at least 14 agree that the items reflect the intended target constructs or that the KSAs assessed are important for successful classroom performance.

5 Results

5.1 Results of the Content Relevance Ratings

Descriptive statistics of the content relevance ratings and the average CVIs for each item type are provided in Table 3. As the table shows, all listening item types had an average CVI above .80. The CVI for the Listening section was .95, clearly above the cut-off criterion. Similarly, all the reading items and item types had a CVI above the cut-off value of .80. The CVI for the Reading section was .95, indicating excellent content relevance.

Table 3 Descriptive statistics and average CVIs for content relevance

Listening item type	Mean	S.D.	CVI
Listen and match	3.66	0.18	0.94
Follow instructions	3.89	0.69	0.97
Question/response	3.45	0.22	0.94
Dialogue	3.48	0.12	0.95
Social-navigational monologue	3.55	0.13	0.93
Narrative set	3.72	0.12	0.94
Academic monologue	3.77	0.07	0.97

Reading item type	Mean	S.D.	CVI
Match picture to word	3.62	0.05	0.89
Match picture to sentence	3.74	0.15	0.95
Sentence clues	3.71	0.13	0.96
Telegraphic sets	3.51	0.14	0.95
Correspondence	3.73	0.11	0.96
Instructional texts	3.74	0.13	0.97
Narrative sets	3.68	0.12	0.93
Expository paragraph	3.79	0.03	1.00

5.2 Results of the Importance of the KSAs Assessed

Descriptive statistics of the importance ratings and the average CVIs for each item type are provided in Table 4. The table shows that six listening item types had an average CVI above .80, with the exception of ‘Academic Monologue.’ The ‘Academic Monologue’ item type is only present in Step 2 of TOEFL Primary. The item type requires test takers to listen to a monologue spoken by a teacher or another adult instructing academic content to students. The test takers then answer three multiple-choice comprehension questions. These questions assess the students’ abilities to understand spoken informational texts and require test takers to have knowledge of organization features of expository texts and the ability to understand key information in a monologue.

A similar degree of agreement among the judges is seen in the Reading section. The majority of the reading item types had a CVI above .80, with the exception of ‘Telegraphic Sets’ that had a borderline CVI of .79. The ‘Telegraphic Sets’ item type is present both in Step 1 and Step 2 of TOEFL Primary. This item type asks test takers to answer multiple-choice questions by locating the relevant information in telegraphic texts in which language is presented in single, phrasal, and short sentence form. Commonly used stimulus materials include posters, menus, schedules, and advertisements. The slightly lower CVI of .79 was considered negligible given that the majority still rated the KSAs assessed in the ‘Telegraphic Sets’ important.

To summarize, the results of the importance of the KSAs assessed by TOEFL Primary indicate high agreement among the judges. The Listening and Reading sections both had an average CVI of .89, suggesting that the majority of the teachers

Table 4 Descriptive statistics and average CVIs for the importance of the KSAs assessed

Listening item type	Mean	S.D.	CVI
Listen and match	3.55	0.22	0.94
Follow instructions	3.55	0.14	0.92
Question/response	3.37	0.18	0.82
Dialogue	3.55	0.07	0.96
Social-navigational monologue	3.61	0.09	0.90
Narrative set	3.70	0.11	0.95
Academic monologue	3.26	0.05	0.72

Reading item type	Mean	S.D.	CVI
Match picture to word	3.69	0.05	0.91
Match picture to sentence	3.76	0.12	0.97
Sentence clues	3.61	0.14	0.92
Telegraphic sets	3.79	0.93	0.79
Correspondence	3.48	0.11	0.84
Instructional texts	3.50	0.09	0.86
Narrative sets	3.68	0.12	0.97
Expository paragraph	3.49	0.07	0.88

considered that the KSAs assessed were important for their respective language teaching contexts.

6 Discussion

This study used CVIs as a research methodology to evaluate the degree of content representativeness of TOEFL Primary. A representative panel of experts was convened to evaluate the degree of match between the test construct and the content of the listening and reading items of the test and to evaluate the importance of the KSAs assessed. The expert teachers' judgments were used as the criterion on which the content-related evidence of validity was based. Results of the study suggest that TOEFL Primary test content largely reflects the target construct being measured and covers the important domains of language knowledge and skills EFL learners are required to possess in order to perform successfully in EFL classrooms.

The content alignment exercise performed by the expert judges identified one listening item type, 'Academic Monologue,' that had slightly lower agreement among the judges, warranting further discussion. As described earlier, the "Academic Monologue" items assess test takers' ability to understand expository texts in a lecture and are more difficult items for the target population. These items were perceived to be less important may be because the listening input was relatively long and for younger learners or lower-proficiency students, the cognitive load of the stimulus materials posed might be overwhelming. It may also be the case that the "Academic Monologue" is designed for learners with higher proficiency level—a level that is higher than the one that the participating teachers were familiar with or currently teaching and thus was considered less important or relevant to their given contexts. Follow-up interviews with the EFL teachers lend a hand to explain the results seen here. One Peruvian teacher, who had 21 years of experience teaching beginner to intermediate English for young learners, indicated that her students had limited exposure to this type of listening input and thought that the academic monologues were too demanding for her students. She said: “We do not have that kind of exercise in the textbook or any other listening task we use in class; we consider this kind of exercise a bit demanding for our students who do not have access to that kind of input neither in their schools nor in their daily lives.”

Other teachers interviewed generally had a positive view about the inclusion of the academic monologues; however, three suggested that the choice of topics should take into consideration young learners’ age and life experience. A French teacher, who had 16 years of experience teaching beginner to intermediate young EFL learners, commented that:

> My students are never exposed to this kind of listening, except when it has to deal with the culture of an English speaking country, such as the life of Nelson Mandela, the religious wars in Ireland, the pilgrim fathers, the constitution in 1776, etc., but not things about insects or for example the earth. Or it would be very general, like not how a volcano works, but the different types of natural catastrophe that you can experience. That is to say, the topic should not be too technical.

This comment indicated that the French teacher’s students, in fact, had exposure to Academic Monologues; however, they were not familiar with the topics included in TOEFL Primary. While this comment highlights the importance of selecting topics that are accessible for young learners who have limited exposure to complex or abstract concepts, it needs to be noted that the teachers’ perceptions of the topic choice might have been influenced by the two academic monologues given to them for evaluation, since both of them were science-related topics. TOEFL Primary encompasses a wide range of topics that represent a variety of disciplines, both in social and natural sciences. The teachers’ views about the topic choice would have been different if different topics had been chosen. Another interesting point worth discussing relates to the French teacher’s remark on introducing topics such as a prominent historical figure from South Africa or the constitution of the United States. These topics, albeit culturally relevant in the French context, may appear to be less familiar for young EFL learners in different parts of the world or EFL contexts.

The teachers’ comments also bring out an important issue in the content design of young learner assessments—topic effects. Whereas the majority of the teachers considered that the Academic Monologue measures what it is intended to measure, the topics of the monologues appear to impact how the teachers perceived the importance of the KSAs assessed with respect to their teaching contexts. This result suggests that there might be a topic effect on the perceived difficulty of task types and potentially on test performance—an effect that can introduce construct-irrelevant variance (Cho & So, 2014). The impact of topics on test performance thus warrants further investigation to inform the choice of topics for the academic monologues.

In terms of research methodology, the investigation suggests that the use of CVIs and an acceptable standard for the CVIs are useful in estimating the degree of content representativeness of newly developed young learner language assessments. On the basis of the results obtained and previous research (Davis, 1992; Lynn, 1986), it appears that content validation of young learner language assessments can be performed by a judiciously selected panel of expert judges who are familiar with the target population and that the experts' judgments can be analyzed using the CVI approach. Emphasis needs to be placed, however, on the careful adoption of a cut-off point that can be used to determine a good degree of content alignment.

7 Limitations and Suggestions for Future Research

A few limitations of the study need to be pointed out. First of all, while the panelists were experienced, representative EFL teachers judiciously selected from varying EFL contexts, the sample size remains small and thus the findings might only apply to the participating teachers' contexts. Future research in validating content representativeness of newly developed young learner language assessments should include expert judges with more diverse nationalities and larger sample size so as to ensure the generalizability of the study results. Secondly, this study evaluated the reading and listening items of the TOEFL Primary test. The computer-delivered speaking test was not included in the evaluation, leaving open the question of the content representativeness of the speaking tasks and the importance of the speaking communication goals for young EFL learners. Subsequent research should investigate the content representativeness of the speaking tasks so that a more comprehensive evaluation of the TOEFL Primary test can be made available to interested EFL teachers and test users. In addition, future research should also investigate whether the mode of test delivery, i.e. paper-based versus computer-delivered, plays a role in how young language learners process input materials and test prompts in order to inform test design. Finally, the study used information from the EFL teachers' judgments of the test items. Other sources of information (e.g., empirical response data) were not available at the time of data collection; however, they should be considered as potential data sources in the future.

8 Conclusion

Results of the study have provided an important piece of empirical evidence to support the content validity of TOEFL Primary and the intended uses of the test. The KSAs assessed by TOEFL Primary listening and reading items were judged to be important and relevant to the content of the different EFL curricula the panelists were familiar with. This finding corroborates with findings from the domain analyses of EFL textbooks conducted in the initial stage of test development and the results of the teacher survey discussed earlier. The multi-stages of test validation have yielded convergent results, consolidating the claims made about the test uses by providing meaningful feedback to support language teaching and learning. In addition, this study presented an evaluative process that can be applied to investigate content representativeness of similar language assessments. Equally important, it suggests a significant role for EFL teachers in the development of new tests for young English language learners.

References

Berk, L. E. (2012). *Child development*. London: Pearson.

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.

Crocker, L., Miller, M. D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*(2), 179–194.

D'Agostino, J., Karpinski, A., & Welsh, M. (2011). A method to examine content domain structures. *International Journal of Testing, 11*, 295–307.

Davis, L. L. (1992). Instrument review: Getting the most from your panel of experts. *Applied Nursing Research, 5*, 194–197.

Fleurquin, F. (2003). Development of a standardized test for young EFL learners. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 1*, 1–23.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247.

Hsieh, C.-N. (2013, September). *Establishing domain representations for a large-scale language assessment for young EFL learners*. Paper presented at the Midwest Association of Language Testers, Michigan State University, East Lansing, MI.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382–385.

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

Nikolov, M. (2016). A framework for young EFL learners' diagnostic assessment: Can do statements and task types. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives*. New York: Springer.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*, 489–497.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5*(4), 285–301.

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

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94–104.

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299–321.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477–481.

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

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

Wu, J., & Lo, H.-Y. (2011). The YLE tests and teaching in the Taiwanese content. *Research Notes, 46*, 2–6.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12*(8), 10–21.