Construct-Irrelevant Factors Influencing Young English as a Foreign Language (EFL) Learners’ Perceptions of Test Task Difficulty

Yeonsuk Cho

Youngsoon So

May 2014
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.
Construct-Irrelevant Factors Influencing Young English as a Foreign Language (EFL) Learners’ Perceptions of Test Task Difficulty

Yeonsuk Cho and Youngsoon So
Educational Testing Service, Princeton, New Jersey

May 2014
Abstract
In this study, we investigated how 12 elementary school children learning English as a foreign language (EFL) interacted with test materials and responded to English language test questions, with the objective of understanding the factors that may unduly influence students’ performance on an English test. The participants were elementary school children in Korea between the ages of 9 and 11. Individual interviews were conducted to elicit young EFL students’ perception of the difficulty of test questions and to gain insight into their thought processes while they were answering listening and reading test questions. Verbal reports indicated that the complexity of language used in test questions may interfere with students’ demonstration of their abilities on the constructs targeted in the test questions. Other issues such as memory load and ambiguity in task descriptions were also observed to influence young test takers’ perception of test difficulty. In the paper, we discuss issues and challenges in designing assessments for young EFL learners.

Key words: cognitive interview, elementary school, English as a foreign language, language test, testing problem, test validity, young children
Assessment of young language learners in English as a foreign language (EFL) contexts is a growing area of practice and research as English instruction is being introduced at elementary or primary schools in many parts of the world. Researchers have recommended that the cognitive development of young language learners be taken into account in test development to ensure valid interpretations of scores (Alderson & Banerjee, 2001; Cameron, 2001; McKay, 2006). However, with some exceptions (Carpenter, Fujii, & Kataoka, 1995; Garcia, 1991; Goh & Taib, 2006; Nikolov, 2006), little research has examined how young test takers interact with tasks on a language test.

Garcia (1991) found that English language learners’ performance on a reading test was negatively influenced by their limited background knowledge and unfamiliarity with the vocabulary used in the test, relative to their English-speaking peers. Carpenter et al. (1995) reported that children often resort to minimal responses, which puts them at a disadvantage in an oral interview, and tend to complete a task even when it is not clear what they are asked to do. Goh and Taib (2006) addressed the metacognition of young second language learners and identified text-related features (e.g., explicitness of information, speech rate) and task-related features (e.g., types of question, types of answer option) that may influence the perceived difficulty of listening comprehension questions. Noting the need for research in the assessment of young language learners, McKay (2006) emphasized the importance of investigating specific issues such as how young learners’ characteristics and task characteristics interact with each other and impact test performance.

The current study was conducted to fill this gap in research by addressing the following research question in an initial stage of the test development context: What are the factors that unduly influence young EFL test takers’ perception of the difficulty of task types and individual items? Data were obtained from cognitive interviews with a convenient sample of 12 elementary school Korean EFL students who answered various types of test questions, with the goal of obtaining practical insights to inform the design of a large-scale international standardized English assessment for young EFL learners.

**Study Design**

**Participants**

Participants were recruited in collaboration with an after-school English program located in one of the affluent areas in Seoul, South Korea. Twelve native Korean-speaking students
between the ages of 9 and 12, with their parents’ consent, participated in the study. When the participants’ teacher was asked to comment on the students’ level of English on a scale of beginner, intermediate, and proficient, the teacher estimated that the participants’ English proficiency ranged from intermediate to highly proficient compared to that of their peers in Korea. It should be noted that even though “highly proficient” was not a point on the scale, the teacher introduced the term to describe some students’ proficiency levels and indicated that their English levels were above “proficient.” More detailed background information about the students was obtained later through individual interviews. Four students reported having spent a significant length of time (from 1 to 2 years) in a country where English is the primary language. Two of these students explained that they moved to Korea when they were 3 to 4 years old, another student stated that he attended kindergarten for 1 year in the United States, and the fourth student said that he lived in the United States for 2 years between ages 5 and 7. Our interview data could not reveal whether those students’ English learning had been impacted by their experiences of living in an English-speaking country during early childhood. However, it is worth noting that there was not a clear relationship between the teacher’s evaluation of the participants’ English proficiency and living in English countries. The four students were not necessarily identified by their teacher as more proficient than those participants who had not had similar experiences; some students who had lived in English speaking countries were evaluated to be less proficient speakers of English than others who had not.

Our participants reported spending a considerable amount of time studying English. In South Korea, formal English instruction begins in Grade 3: Students in Grades 3 and 4 receive a 40-minute class a week, and those in Grades 5 and 6 receive two classes a week (Ministry of Education, Science, and Technology, 2008). However, a majority of the students in our sample, nine out of 12 students, said that they started learning English before formal English instruction began at school. On average, the students in the study sample reported studying English for 10 hours per week; individual estimates ranged from 4 to 19 hours.
<table>
<thead>
<tr>
<th>Skill</th>
<th>Task type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Listening</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Phonemic awareness</td>
<td>Students listen to three words (e.g., <em>chair, share, chair</em>) and choose the one that is different.</td>
<td></td>
</tr>
<tr>
<td>2. Print to sound</td>
<td>Students match a written word to one of the three words they heard.</td>
<td></td>
</tr>
<tr>
<td>3. Sound to print</td>
<td>Students listen to one word and match it to one of the three written words.</td>
<td></td>
</tr>
<tr>
<td>4. Sentence-picture matching</td>
<td>Students listen to one short sentence and match it to one of the three pictures.</td>
<td></td>
</tr>
<tr>
<td>5. Directions-picture matching</td>
<td>Students listen to directions and match them to one of the three pictures.</td>
<td></td>
</tr>
<tr>
<td>6. Question-answer matching</td>
<td>Students listen to a short question or statement and identify an appropriate rejoinder from three rejoinder options.</td>
<td></td>
</tr>
<tr>
<td>7. Conversation comprehension</td>
<td>Students listen to a short conversation and answer one comprehension question. The question and options are both spoken and written.</td>
<td></td>
</tr>
<tr>
<td>8. Story comprehension</td>
<td>Students listen to a story and answer multiple comprehension questions. The question and options are both spoken and written.</td>
<td></td>
</tr>
<tr>
<td><strong>Reading</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Word-picture matching</td>
<td>Students read one word and match it to one of the three pictures.</td>
<td></td>
</tr>
<tr>
<td>2. Sentence-picture matching</td>
<td>Students read a short sentence and match it to one of the three pictures.</td>
<td></td>
</tr>
<tr>
<td>3. Form and meaning</td>
<td>Students fill in a gapped sentence with one of three words provided (e.g., <em>they, them, their</em>).</td>
<td></td>
</tr>
<tr>
<td>4. Story comprehension</td>
<td>Students read a story and answer multiple comprehension questions.</td>
<td></td>
</tr>
</tbody>
</table>
Instruments

The two instruments in this study included multiple-choice questions divided into listening and reading sections to measure students’ receptive English skills and an interview protocol. The multiple-choice questions were newly developed questions (the so-called prototypes) proposed for an international English test for young EFL learners. These test questions were developed to measure a set of receptive English skills that commonly appear in a number of elementary school EFL curricula. Eight task types were proposed for the listening section and four for the reading section. The complexity of these tasks varied considerably, ranging from simple phonemic awareness in listening and word-level comprehension in reading to extended story comprehension. More information about task types is presented in Table 1. Each of the two sections included 50 three-option multiple-choice questions. The listening section was administered through a CD.

Interview questions were developed to collect participants’ background information and to elicit their perception of the difficulty of test questions and the thought processes they followed in selecting the answers to the test questions. The following questions were used to prompt the participants to comment on their experience with the test:

- Were the directions clear and easy to follow? If not, why?
- Which question(s) was (were) the easiest? Why?
- Which question(s) was (were) the most difficult? Why?
- How did you figure out the answer?

Student responses to the above questions were then used to build the next questions to probe further and clarify the initial responses.

Procedures

The data were collected through individual 1-hour study sessions that included the test administration and the interview. Each student was administered one of the prototype test sections: six students took the listening section, while the other six took the reading section. Students were assigned to either group (i.e., listening or reading) according to their age and reported proficiency levels so that the two groups were similar with respect to the two variables.

The sessions took place in a classroom at the participants’ after-school English program, and interviews were conducted by one of the authors in Korean except for one student, who
wished to do the interview in English. This student did not live in an English-speaking country, and was considered to have an intermediate level of English proficiency by her teacher. The sessions were audio recorded and later translated and transcribed. Observation notes were also taken during each session.

During each session, the purpose and procedure of the study were explained to the student, after which the background questions were asked. Students were told that they would be asked to comment on their test-taking experience after answering test questions. The subsequent procedure varied depending on the section to which students were assigned. The reading section was self-paced within the allotted 30 minutes. The time students needed to complete the reading section ranged between 14 and 28 minutes, with the average being about 19 minutes. The interview questions were asked after the student had finished the reading section. This procedure was modified for the listening section, where the test administration was guided by the recordings on the CD. Because it was difficult for the students to remember the details of all the test items after they had completed the listening section, the recording was paused after every two or three test questions in order to ask the interview questions. This procedure allowed the students to articulate their thought processes while remembering the content of test questions that they had just answered. The recording was replayed during interviews if it was deemed necessary to help the students clarify and fill in the gaps in their explanations.

**Interview Data Coding and Analysis**

Because the study was exploratory, we did not impose an a priori coding scheme on the data. Instead, the two authors independently reviewed the interview data (both recorded interviews and transcripts), marked parts of the interview that seemed relevant to our inquiries, and coded them according to emerging themes. After coding each interview independently, the authors met to discuss the coding results and to reconcile discrepancies. To summarize the data, we tallied the number of students whose comments were coded for a particular theme (or coding category) and subsequently combined results across the six participants separately for reading and listening.
Results
The interviews with the students revealed a number of construct-irrelevant factors influencing young language learners’ perception of the difficulty of test questions. The categories that emerged from the analyses of students’ comments and behaviors are presented in Table 2. The figures in the table represent the numbers of students whose comments were relevant to the corresponding categories. Interestingly, more issues were observed with regard to the listening tasks, although some of these issues (e.g., ambiguity in task description) could have been possible for the reading tasks as well. The factors that could not have been relevant to the reading section (e.g., speaker’s voice) are denoted as “not applicable” in the table.

The complexity of language in questions and options was the most predominant issue with both the listening and reading tasks. Four of the students taking the listening section and five of those completing the reading section reported some degree of difficulty understanding a test question, answer option, or both. The interview excerpt in the following example demonstrates how sometimes the linguistic complexity of a test question and options hinders young EFL students’ test performance when understanding the wording of a test question and options is not part of the target construct of a test item. The test question in Figure 1 was a conversation comprehension question (Task 7).

Table 2
Construct-Irrelevant Factors Related to Task Difficulty

<table>
<thead>
<tr>
<th>Coding category</th>
<th>Listening (n = 6)</th>
<th>Reading (n = 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity of language in questions and options</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Difficulty remembering information</td>
<td>4</td>
<td>n/a</td>
</tr>
<tr>
<td>Ambiguity in task description</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Child speaker’s gender unclear from voice</td>
<td>3</td>
<td>n/a</td>
</tr>
<tr>
<td>Simultaneous processing of listening and reading</td>
<td>2</td>
<td>n/a</td>
</tr>
<tr>
<td>Ambiguity in pictorial options</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. n/a = not applicable.
Question 34: What surprised the girl?

A. Who made the food
B. Where the boy made the food
C. How much food the boy made

Figure 1. A conversation comprehension question.

Ella: Question 34 was a bit confusing. Between B and C.

Interviewer: Why? A was OK?

Ella: Well, because the speaker said that the boy made the food, and B and C say so.

Interviewer: Why did you think B was the answer?

Ella: Hmm . . . just.

Interviewer: Why were you sure that A was not the answer?

Ella: Because the speaker said that the boy made the food, but A is asking who made the food.

The student’s comments on the item suggested that she did not understand the meaning of the options because of their grammatical resemblance to that of interrogatives. This tendency to associate wh- words with interrogatives was also observed with another item. One student, Katie, commented that “if we use why or what, it has to be a question” when describing how she chose a correct answer from among the three options (beginning with why, when, and what, respectively) for a form and meaning item (Reading Task 3).

The next most frequently mentioned factor concerned the amount of information that test takers needed to remember during the listening tasks. Four of the six students who took the listening section commented that they had difficulty remembering what they had heard. This comment was particularly common with Listening Tasks 6 and 8, which required test takers to hold more information in their memory compared to the other listening tasks. For instance, on the question-answer matching task (Task 6), test takers had to remember a spoken question or statement followed by three spoken options. This particular item format seemed to have strained
young test takers’ working memory capacity, which is not the direct target of assessment, as illustrated in the following example.

Lisa: Question 17. I suddenly forgot the question.
Interviewer: What do you mean? You seemed to be paying attention.
Lisa: I was paying attention but when I heard Option A, the question I remembered just disappeared.
Interviewer: How did you answer 17?
Lisa: Just a guess.

A similar comment was made by another student about the story comprehension task (Task 8), in which a long text was followed by a set of questions.

Sam: This [referring to one of the stories] also required memory. I had to listen to one story and answer all these questions. The story was long. There were many things that happened, and you can’t remember everything. Because I don’t have good memory, it was difficult. Not because I didn’t understand the story. It was difficult because I couldn’t remember everything.

The lack of clarity in the task description was another factor observed that could have negatively influenced young students’ test performance. The directions-picture matching item shown in Figure 2 was designed to measure the ability to follow spoken directions. In Figure 2, the test takers first listened to a mother telling her daughter to wake up and then were asked to answer “Which picture shows what the girl should do?” by choosing a picture. However, the purpose of this task was not clearly understood by some test takers, as illustrated in the following excerpt.

Irene: I didn’t know which part was the question or directions. On this one [pointing to the test item], I was not sure whether the mom had to say so [referring to “wake up”] because the girl was sleeping or whether the girl got up because of what her mom said.
Interviewer: What did the question ask?

Irene: Hmm . . .

Interviewer: Why did you choose B?

Irene: Mom said to wake up, so I chose B.

**Figure 2. A directions-picture matching item.**

Although the student in this example answered the test item correctly, she first commented that she could not distinguish the test question (i.e., mother’s directions) from the task description. Furthermore, the student’s account of her thought process suggested that it was not clear whether she was supposed to find a picture depicting the cause (Option A) or outcome (Option B) of what the mother told her daughter to do.

The interview data also revealed that some well-intended task features could be sources of confusion for young test takers. For example, the prototyping tasks were designed to improve young test takers’ task engagement by making use of children’s voices and many visual images; some of the conversations included in the present prototyping study were dialogues between a boy and a girl. Success on certain comprehension questions hinged upon whether test takers were able to differentiate the gender of each speaker, because the question was phrased, for instance, as “What did the boy want to do?” However, some students reported that they could not identify the gender of young children’s voices in conversation.

Another task feature that was found to negatively influence some students’ perception of task difficulty was the simultaneous presentation of the test questions and options in oral and written formats (i.e., both voiced on the CD and printed in the test booklet) for Tasks 7 and 8.
This was intended to minimize memory load, but it introduced another problem by diverting students’ attention, as the following excerpt indicates.

Sam: Because when [the speakers] talk, this, this, and this [pointing to the written question and options], these are mentioned, right? If I listen to the conversation without reading the questions and options, I don’t know what it is. So I keep reading them, then I miss what they have said. So it’s difficult to answer the question.

**Discussion and Conclusion**

Language assessment research and practice, in general, have come a long way toward understanding the characteristics of test takers and the characteristics of test tasks (Cohen, 2012) and drawing attention to the influences of such factors on test performance. Such understanding is of significance because construct-irrelevant factors may undermine the interpretation of the test results. However, there has been little research to demonstrate how young EFL test takers’ test performance is influenced by test task characteristics. We investigated how a small group of elementary school EFL students attributed task difficulty in order to identify unforeseen issues with newly developed test questions and task types.

Before discussing the findings, we would like to acknowledge the limitations of the study. The study was conducted using a small convenient sample of 12 nine to eleven year old elementary school students, who may not have been representative of the majority of students learning English in Korea or in other countries. As described early in this paper, all the students in the study had advanced English skills, according to their teacher. Many students reported having spent a considerable amount of time learning English, and some had lived in an English-speaking country. It is thus important to realize that the findings presented in the study are only suggestive in nature and that further research is needed. The interview procedure also limits the ability to generalize the results beyond the context of this study. The intrusive nature of the interview procedure, in which participants were asked to comment on listening questions during the test, could have influenced participants’ performance and articulation of thought processes in answering subsequent test questions. Another issue that could not be clarified in the study was that, while several factors related to the perceived difficulty of listening questions were observed, very little was revealed related to reading questions (see Table 2). At least two explanations are
possible. First, the two groups of students assigned to reading and listening sections may have differed with respect to the ability to articulate their thoughts. Moreover, the different interview procedures might have led to more comments being made on the listening tasks than the reading tasks. The interview procedure used for the listening section allowed the students to reflect on their thought processes after two to three test questions, whereas for the reading section, the interview questions were not asked until the end of the test.

In spite of these limitations, the study findings show how a small group of elementary school EFL students interacted with test questions and provide valuable feedback to test developers. As described earlier in this paper, this study was conducted as part of test development efforts to understand how various aspects of tasks could have unintended effects on test performance. The findings of the present study prompted revisions of the prototype test questions as follows.

First, the fact that the most frequently reported difficulty was related to language complexity highlights the importance of a general item-writing principle: the test questions and task directions must be kept simple and clear. Although this principle applies to test development for any age group of test takers, it should be emphasized even more when developing test questions for young test takers. Unfortunately, we have little information as to what aspects of language may pose challenges to young language learners; for example, what are the grammatical structures and vocabulary that should be avoided for young EFL learners? In the current context, item specifications were revised to eliminate the use of the grammatical structures discussed in the results section.

Second, and on a related note, our study shows the importance of clear task descriptions, especially for young EFL learners. Our finding is in line with observations in the literature indicating that, possibly for an emotional reason, children are less likely than adults to seek clarifications when encountering unclear tasks (Cameron, 2001; Carpenter et al., 1995). Probably for a similar reason, some students in our study reported that they answered the test questions even though they were unsure what they were being asked to do. To clarify the task expectations for young learners, it was decided to revise some of the task descriptions and provide more sample items where necessary.
Next, the amount of information to remember was also frequently mentioned as a concern by the students who completed the listening section. In second-language listening comprehension research, language learners’ working memory is considered relevant to their listening ability (Goh, 2002). However, the role of memory in listening comprehension is not well understood when it comes to the assessment of young language learners because at this young age, working memory and other cognitive functions are not fully developed. Based on this finding, changes were made to two listening task types. For the story comprehension task, the length of passages was shortened. On the question-answer matching task, some students stated that it was challenging to remember a question that was spoken only once; therefore, the task was revised so that the question is repeated with every option. In addition, some students reported having difficulty differentiating the voices of the boy and the girl in the conversation comprehension task. This unanticipated difficulty provided specific insight into the importance of selecting voices for dialogue items so that construct-irrelevant variance is minimized. Our findings showed that the gender of children’s voice before puberty is not easily discernible. As a result, older children are now used to record conversations used in the test.

Finally, the study indicates that young EFL learners may find it challenging to read and listen at the same time for a particular item. This difficulty with a task characteristic that is usually helpful for older test takers may be related to a “literacy skills lag” (Cameron, 2003, p. 108) in child language development. When students have not acquired sufficient literacy to read written questions and options with ease, much of their cognitive capacity may be devoted to the task of reading, leaving a limited capacity to simultaneously pay attention to spoken language. Presumably, the students in the study have more developed English skills than their peer students. Thus, the impact of literacy skills on a listening test could be of far greater concern than suggested in the present study and should be carefully considered in designing a test for young students. However, it can be argued that providing written questions and options may help lessen young test takers’ memory load by serving as an anchor for the information they need to remember during a listening test. For this reason, it was decided to provide questions and options both aurally and in print while simplifying the language. This result also implies that instruction on test-taking strategies, particularly instructing test takers to focus on listening while listening passages are played, may help young EFL test takers navigate the task demands during testing (Goh & Taib, 2006) with greater ease.
The present study has addressed important but understudied areas, including characteristics of young language learners, task characteristics, and the interaction between the two. We hope that our study findings provide practical and concrete design suggestions for developing assessment tasks for young language learners.
References


Notes

1 All participants’ names are pseudonyms.