



CaMLA Working Papers

2015-03

A Validation Study of the Reading Section of the Young Learners Tests of English (YLTE)

**Paula Winke, Shinhye Lee, Irene Jieun Ahn,
Ina Choi, Yaqiong Cui, Hyung-Jo Yoon**
Michigan State University
United States





A Validation Study of the Reading Section of the Young Learners Tests of English (YLTE)

Authors

**Paula Winke, Shinhye Lee,
Irene Jieun Ahn, Ina Choi,
Yaqiong Cui, Hyung-Jo Yoon**
Michigan State University

About the Authors

Paula Winke received her PhD in applied linguistics from Georgetown University in Washington DC. She is currently an associate professor in the Department of Linguistics and Languages at Michigan State University. There she directs the Master of Arts in Foreign Language Teaching program, and teaches in the Second Language Studies and TESOL graduate programs. She researches language assessment practices, language teaching methods, and task-based teaching and assessment. She is the 2012 recipient of the International TESOL Association's Distinguished Researcher Award.

Shinhye Lee received her BA in history education and english education from Hongik University (Korea) in 2010 and MA in english education from Ewha Womans University (Korea) in 2012. Since joining the Second Language Studies program at Michigan State University in 2013, she has been working as a research assistant to Dr. Paula Winke for language assessment and for MSU's Language Proficiency Flagship initiative. Her research interests include language assessment and test practices, speech production, and young language learners.

Jieun (Irene) Ahn received a BA in English language and literature and a MA in teaching Korean as a foreign language from Ewha Womans University in Korea. She is currently a PhD student in Second Language Studies at

Michigan State University. Her research interests revolve around psycholinguistic approaches to second language acquisition, interactionist approaches to instructed second language learning, and the use of eye-tracking methodology in second language research. She was awarded a full research assistantship for her PhD studies and has been working as a research assistant for Drs. Susan Gass, Paula Winke, and Aline Godfroid.

Ina Choi holds an MA in foreign and second language education from the Ohio State University. As a doctoral student in Second Language Studies at Michigan State University, she is currently working on survey development and data analysis for the Center for Applied Inclusive Teaching and Learning in Arts and Humanities (CAITLAH). She has also taught third-year Korean and pedagogical grammar at MSU. Her research interests include second language vocabulary acquisition, incidental vocabulary learning and online processing, and task-based teaching and learning.

Yaqiong Cui received her BA in teaching Chinese as a second language at East China Normal University (China)

and a MA in East Asian languages and cultures at the University of Illinois at Urbana-Champaign where she also taught Chinese at various levels. Since joining the Second Language Studies program at Michigan State University in 2012, she has been working as a research assistant and language facilitator at the Center for Language Teaching Advancement. Her research interests involve psycholinguistic approaches to second language acquisition and processing (particularly Chinese), second language reading development, and language assessment.

Hyung-Jo Yoon earned his BA in english education from Hankuk University of Foreign Studies (Korea) in 2009 and his MEd in TESOL from the University of Pennsylvania in 2013. He is currently a PhD student in Second Language Studies at Michigan State University. While pursuing his PhD, he has been working as an English instructor (teaching reading/writing and composition courses) at the English Language Center. His research interests include second language writing, language assessment, and computational analysis of natural language.

Table of Contents

Abstract.....	1
Test Validation: How It Can Be Done	2
Why We Investigated the Validity of the YLTE.....	3
Methodology.....	5
Results and Discussion.....	8
Qualitative Results	9
Conclusion.....	20
References	22
Acknowledgements.....	24
Appendix	25

Abstract

In this study we investigated the validity of the reading and writing sections of CaMLA's Bronze and Silver Young Learners Tests of English (YLTE). A test's validity can be analyzed from many angles. We took the following approach: First, we evaluated whether the tests are appropriate for measuring the reading and writing skills of a particular group of learners: 19 English language learners (ELLs) ages 7 to 9. We also looked specifically at the *cognitive validity* (Weir, 2005) of the tests, that is, whether the tests measure the skills intended by the test developers. We followed Green's (2014) suggestions for monitoring a test's cognitive validity: (a) we observed how the children performed and analyzed (qualitatively) their test-taking behaviors, and (b) we interviewed the children to try to understand what they thought about the test, how they found a correct answer, or how they decided on their responses.

Seven native speakers and 12 ELLs (with Korean or Mandarin Chinese native languages) took the tests. We videotaped the children as they took the tests, had each draw a picture of how he or she felt during each test, and interviewed the children about their test-taking experiences. Given the score outcomes, the tests appear reliable and consistent in discriminating learners from native speakers. Analyses indicated that three items on the Bronze test (out of 25 items) and five on the Silver (out of 40) were more difficult for native speakers than for ELLs. We showcase those eight items and use our qualitative data and research into child language development to propose reasons why the items were inversely discriminating. We argue that piloting on native speakers can reveal when incorrect responses stem from something other than reading or writing problems, such as from a lack of assessment literacy, developmentally-appropriate overgeneralizations of grammatical rules, or age-related limitations in morphological-rule learning or cognitive control. We conclude that all tests can be improved, even those that are already structurally and psychometrically reliable and valid.

English-language literacy is vital for English language learning (Grabe, 2009) because learners glean new vocabulary, grammar, and cultural information when they read (Laufer, 2003; Pulido, 2004; Webb, 2005). This is especially true for children. With good reading skills and access to age-appropriate literature, children can continue learning outside the classroom (Dewey, 2004; Paribakht & Wesche, 1999). Such learning via reading is essential for children, especially for those growing up in homes in which English is not the main language used by the family. According to the United States National Center for Education Statistics (National Education Association, n.d.), children who read frequently develop stronger reading skills and have higher overall success in school. They also progress further in school, attend institutions of higher education at a higher rate, and go into STEM (science, technology, engineering, and math) fields at higher rates.

Concomitantly with educational emphasis on reading, there is increased demand for reliable and

valid reading assessments to measure the success of reading programs. Such assessments are used to provide diagnostics concerning individual children. Reading-test results can be used to inform teachers and parents of the reading strengths and weaknesses a child may have and guide any individual, educational reading plans for that child. In addition, the results can be used to evaluate the outcomes of the reading program at the school.

Ideally, variations in English language reading test scores should be attributable to the children's English-language reading skills alone, which would indicate the test measures English-language-reading skills as it should; a test should not have a significant portion of the test-score variance attributable to other factors beyond reading, which would be considered *construct irrelevant variance*. Unfortunately, as testing specialists know, all tests have measurement error, especially when children are involved (Biggar, 2005): children do not always try their best (Hasselgreen, 2000; McKay, 2006); they get distracted, tired, anxious, or bored; some don't

understand the directions, and test administrators may not be allowed to explain, which can be confusing and stressful for children (Menken, 2008); sometimes children just make patterns on the optical answer forms, commonly known as “bubble sheets” (Winke, 2011). Thus, ensuring that a reading test for children produces reliable and valid scores requires more than a quantitative evaluation of the test’s outcomes. Robust qualitative analyses are needed to understand why certain tasks might potentially be uninformative or developmentally inappropriate for certain children. In this study, we investigate young children’s test-taking processes and explore their test-related opinions and reflections to shed light on the validity of the CaMLA (Cambridge Michigan Language Assessments) Young Learners Tests of English (YLTE). We examine the validity of the test to better understand whether the test appropriately measures what it is supposed to, and whether the test is justified in terms of its outcomes, uses, and consequences (Bachman, 1990; Hughes, 2003; Messick, 1989).

Test Validation: How It Can Be Done

In educational measurement, researchers have long debated exactly what test validity is and how to evaluate whether a test is valid. We define validity using a quote from Henning (1987, emphasis original), as presented by Alderson, Clapham, and Wall (1995, p. 170): “A test is said to be valid to the extent that it measures what it is supposed to measure. It follows that the term *valid* when used to describe a test should usually be accompanied by the preposition *for*.” Alderson et al. noted that if a test is used, the validity of that *use* needs to be established and demonstrated. The questions they suggested test-score users ask are “How do you know this test is valid?” and “For what purposes is this test valid?” Even when a test’s validity is explained to the test-score users (typically by the test developers in a validation or reliability report), Alderson et al. recommended that the test-score users still use their own judgement to decide, based on the evidence provided, whether the test is valid, or to what degree it is valid. This is because validity is not an absolute and must be contextualized to include both test takers and test purposes.

Researchers have described different *types* of validity (Alderson et al., 1995; Chapelle, 1999; Green, 2014; Norris, 2008) that are often used to show a test is valid. Each provides unique information from different vantage points. We describe here the four that Norris (2008) suggested are the stereotypical *types* of validity that test

developers often fall back on as the bare basics of test validity research.

1. *Content validity* is whether a test represents (or samples from) the content (the skill being assessed) well enough. For example, does a test of reading comprise a well-balanced sample of the different types of texts and genres that one would expect the test takers to know? Content evaluation typically involves asking experts (subject teachers/specialists) their opinions on whether they feel the test measures all the different areas or varieties of the skill it should. Those evaluating for content validity can also compare the test’s actual content against a list of what the content ought to be (e.g., the text types or genres taught to prospective test takers).
2. *Concurrent validity* is whether the test scores align with some other, trusted external measure of the same underlying skill. For example, does a teacher find that her best readers do best on the reading test? Do the test scores correlate with another, long-standing measure of the same skill? Concurrent validity is perhaps the easiest validity type for test developers to present because it can be expressed in terms of a correlation coefficient or level of agreement between two sets of scores.
3. *Predictive validity* is whether the test scores align with some future measure of the same underlying skill. That is, does a high reading score accurately predict high reading performance in the real world? This type of validity evidence makes sense in certain contexts. For example, if a reading test is being given for placement purposes, one would expect the test to predict who would do well in a beginning class and who would do well in an intermediate class. The test would lack predictive validity if some placed into the intermediate class failed.
4. *Construct validity* is, according to Alderson et al. (1995), the most difficult part of validity to explain. This is because it brings to question whether a test is measuring what it is supposed to be measuring. Norris (2008) suggested that investigating a test’s construct validity is most often done for tests that purport to measure some type of psychological state, like aptitude, anxiety, or motivation. Like content validity,

construct validity is often evaluated through expert judgement. Experts can inspect the test and attest to whether it measures what it is purported to measure.

Norris (2008) explained that over the decades, language test developers have used a selection of these four validity measures (content, concurrent, predictive, and construct validity) to establish the validity of tests after pilot testing but before large-scale test use. As such, validity is often reported as a test characteristic and “a quality of the test instruments rather than the interpretations based on test scores and the uses to which they were put” (Norris, 2008, p. 38). Norris pointed out that more is needed and that validity should be interpreted with test-score uses in mind. Validity evidence should be gathered not just for reporting purposes, but also for making the existing test better. Bachman and Palmer (2010) called this expanded notion of validity a process of *assessment justification*, which is when researchers (1) gather evidence that the test score uses are justified and (2) explain those justifications to stakeholders. In testing programs, assessment justifications should be ongoing. Validation study results should feed back into the testing system, making it better (Chapelle, 1999; Norris, 2008). Indeed, Messick (1989) summarized changes involved in validity estimation when he writes that empirical evidence and theory have to converge to show that test scores allow for *appropriate inferences* and *actions* (p. 13). Messick wrote that validity evidence can and should include (a) analyses of the ways in which test takers respond to test tasks, and (b) investigations of test processes across groups.

More recently, Messick’s (1989) call to better understand test takers’ response processes has been coined as investigating the *cognitive validity* (Weir, 2005) of a test. Green (2014) explained that “one approach to cognitive validation is to ask the assessee about how they carry out tasks” (p. 81). He suggested this can be done through verbal protocol methods that evaluate how test takers found a correct answer or how they decided to respond. Green stated it is also possible to analyze test takers’ test-taking behaviors to understand better how they perform their test tasks. In this study, we aim to do these things. We aim to investigate the cognitive validity of reading tests for children to see if the tests are valid in measuring the reading skills of the children who take the tests. We do this not only to validate the tests, but also to provide feedback to the test developers so that changes to the test, if needed, can be made.

Why We Investigated the Validity of the YLTE

We became interested in the validity of foreign and second language tests for children after Winke (the lead author on this study) investigated the validity of a large-scale, high stakes test for K–12 English language learners in the state of Michigan (Winke, 2011). In that study, Winke investigated teachers’ opinions of the Michigan English Language Proficiency Assessment (ELPA) after the teachers administered the assessment to children ages 5 to 18. The test included reading, writing, listening, and speaking sections. Winke surveyed 267 teachers about the exam process. The teachers noted that the testing had both positive and negative consequences. The mandated tests made their programs more visible (a positive consequence), but results were not valid for all test takers because the test was too difficult for some, which made certain children feel badly about themselves (a negative consequence). Teachers suggested the test scores were possibly uninformative for very young test takers (ages 5 to 7) because their scores may have had much construct-irrelevant variance. The teachers noted young children could not concentrate during the test or were unwilling to participate in some parts of the test. For example, students sometimes had to talk to strangers to take the speaking test, and the young students were reluctant to talk to someone they did not know, which resulted in lower-than-expected speaking-test scores.

Winke (2011) demonstrated that collecting qualitative data from stakeholders provides rich and important information about the broad validity of a testing program. Such information can be used to ensure that large-scale testing programs for children are accountable not only to the entities that mandate them, but also to those the tests intend to serve, the stakeholders. Winke noted that her study was limited in scope because only the teachers were surveyed. The broad validity of a test program could be more thoroughly evaluated through qualitative data from other stakeholders, including the test takers.

Following Winke (2011), we asked: What additional validity evidence would observations of and interviews with child test takers provide? Might qualitative data from children help researchers better understand children’s testing processes? Can researchers obtain from children the same type of cognitive validity evidence that other researchers (i.e., Field, 2009) have found with adults?

Carless and Lam (2014) investigated lower elementary school children’s perceptions of their school-

based testing experiences in Hong Kong schools. The children were predominately 8 years old and in the third year of primary school. Carless and Lam adapted data collection methods originally developed by Hall, Collins, Benjamin, Nind, and Sheehy (2004) and Wheelock, Bebell, and Haney (2000), who employed focus groups and picture-drawing tasks (respectively) to investigate young children's perspectives on large-scale testing programs in Britain (Hall et al.) and in the United States (Wheelock et al.). Carless and Lam were not investigating language-learning exams per se; rather, they were investigating the culture of examinations and how they affected children.

Carless and Lam had 115 children participate in 21 focus group sessions, during which they asked the children about their perceptions of tests; 76 children drew pictures of their test taking experiences. The researchers coded the drawings and the focus group data as positive, negative, neutral, or mixed, and they also coded along three themes: affective response to assessment, parental influence, and connections between testing and learning. In sum, the authors found negative perceptions slightly outweighed positive ones. They suggested that these data, from children in the early years of schooling, were concerning. They wrote that the data "represents for us a cause for concern in that over the longer term they [tests] may impact negatively on students' willingness to engage fully and productively with school life" (p. 321).

In this study, we conduct a small-scale investigation into the validity of the reading- and writing test items that appear in the CaMLA Young Learners Tests of English (YLTE) Bronze and Silver tests. In particular, we research whether these two tests are developmentally and contextually appropriate for English-language-learners at ages 7, 8, and 9, the youngest population for which these tests are targeted. We do this to add to the validity arguments presented in the 2014 YLTE Report (CaMLA, 2014) and to investigate further our assumption that the process of qualitative *assessment justification* will work and prove beneficial in the context of child language assessment. We intend to investigate the cognitive validity of the tests by following Messick's (1989) recommendations.

In this study we investigate two groups of children: native- English-speaking children and children learning English as a second language. By employing these two groups of children, we hope to be able to pinpoint the sources of any difficulties or wrong answers. If specific difficulties or wrong answers surface in the nonnative-

speaker group only, then we may be able to attribute them to a lack of English (which is what we would expect). If the difficulties or wrong answers appear across both groups, then we will speculate about their sources. Most likely, in such cases we would look for construct-irrelevant sources, such as cognitive validity issues, a lack of cultural background knowledge, or misconstrued test directions or tasks.

The YLTE tests are an excellent context for this study because the tests have already proven to be valid in a number of ways. The YLTE program actually comprises a suite of three tests—Bronze, Silver, and Gold—which refer to the proficiency levels at which these tests are targeted. In this study we focus on the first two (Bronze and Silver), mainly because we have limited resources, but also we want a small but in-depth investigation. Complete descriptions of the tests are available on the CaMLA website at CambridgeMichigan.org. We note that CaMLA, in collaboration with Cambridge English, takes great care in designing the YLTE tests. They ensure that the tests have excellent content validity, are fun and motivating, and provide a clear and transparent assessment of young learners' English skills. While the YLTE 2014 Report does not provide any concurrent validity evidence or predictive validity evidence, it does allude to them. The report states the following:

CaMLA is committed to the excellence of its tests, which are developed in accordance with the highest standards in educational measurement. All parts of the examination are written following specified guidelines, and items are pretested to ensure that they function properly (p. 1).

The Bronze and Silver tests assess reading comprehension through a variety of text types and item formats. The Bronze test was designed to be easier than the Silver test and has fewer sections and questions. The texts across both tests range from simple noun phrases to three-paragraph stories. The item types include multiple-choice, true/false, and one- to three-word completions, some with word banks, others without. Thus, the tests assess reading *and* writing, in that children are required to interact with given texts by supplying written responses. The final score reported for the children is a single, reading/writing composite score. The answer key further requires test takers to have accuracy in spelling and grammar. In this study, we used the sample Bronze and Silver tests available online and the answer keys

CaMLA provided. We used these two tests to explore the following research questions:

1. Do English-language-learning children (ages 7, 8, and 9) perform better on the YLTE Bronze test than on the Silver test (i.e., in line with the progression expected by the test developer)?
2. Are native English-speaking children (ages 7, 8, and 9) able to perform better on the Bronze and Silver YLTE tests (as they would be expected to) than the same-aged, nonnative-speaking, learners of English?
3. Do children lose attention during the tests, and if so, why?

Methodology

Participants

Nineteen children participated in this study: 12 nonnative speakers of English (six Mandarin Chinese-speaking children and six Korean-speaking children) and seven same-aged, native speakers of English. Fifteen were girls and four were boys. The nonnative speakers had been learning English for an average of 23 months (SD = 16), with the range from one month (Participant 7) to 46 months (Participant 12). The nonnative speakers had been living in the United States for an average of 5 months (SD = 3), with the range from 1 (Participant 7) to 10 months (Participants 13 & 14). A list of the 19 participants, ordered by their composite (Bronze and Silver) scores (from high to low), is in Table 1.

Table 1: Participants sorted by their composite (Bronze and Silver) test score (descending), and then by ID# (ascending)

ID #	Age at Test	Gender	L1	Months Learning English	LoR (Months) in USA	School Grade	Parent Ed	Books Per Day	Book Language	Bronze Test Score	Silver Test Score	Total Score
5	9	F	English	NA	NA	4	5	1	Eng.	24	36	60
11	9	M	Chinese	4	5	4	6	4	Ch.& Eng.	24	36	60
19	7	F	English	NA	NA	2	6	5	Eng.	22	38	60
6	9	F	English	NA	NA	3	4	1	Eng.	25	34	59
1	7	F	English	NA	NA	2	5	1	Eng.	22	35	57
2	8	F	English	NA	NA	2	5	1	Eng.	21	35	56
3	9	F	English	NA	NA	3	4	1	Eng.	20	35	55
18	9	F	Korean	37	3	3	4	3	Eng. & Kor.	23	32	55
17	9	M	Korean	36	3	4	5	1	Eng. & Kor.	23	30	53
4	8	F	English	NA	NA	2	4	2	Eng.	20	31	51
9	8	M	Chinese	12	7	2	5	2	Ch. & Eng.	22	29	51
10	8	F	Chinese	42	5	2	4	1	Eng. & Ch.	23	27	50
8	7	F	Chinese	36	6	2	6	1	Ch.	20	29	49
15	8	F	Korean	12	8	3	4	3	Kor. & Eng.	23	25	48
13	7	F	Korean	18	10	1	4	3	Eng. & Kor.	18	23	41
16	7	F	Korean	12	3	2	5	1	Eng. & Kor.	15	22	37
12	9	F	Chinese	46	2	4	6	1	Ch. & Eng.	12	17	29
14	7	M	Korean	18	10	1	5	2	Eng. & Kor.	17	12	29
7	7	F	Chinese	1	1	1	5	1	Ch.	7	NA	7

Notes: Parent Ed = highest education level met by a parent, with 6 = PhD; 5 = MA, MS, or JD, & 4 = BA or BS.
Books Per Day = # of books on average the child reads or has read to him/her.

Materials

Background questionnaire. For this study we designed a one-page background questionnaire, which was presented in the native language of the parents, who were allowed to respond in their native language. Information from the questionnaire is in Table 1.

Reading and writing tests. The CaMLA Bronze and Silver sample tests we used are available on CaMLA's website (<http://www.cambridgemichigan.org>). The Bronze test of reading and writing takes approximately 20 minutes to complete. It has 5 parts with 25 questions total. The Silver reading and writing test takes approximately 30 minutes to complete and has 6 parts with 40 questions total.

Child drawings. In this study we used the draw-a-picture technique (Carless, 2012; Carless & Lam, 2014; Wheelock et al., 2000). The script used to explain the picture-drawing task to the child follows (this was translated into Chinese for the Chinese children and into Korean for the Korean children):

Thank you for taking the test. Now please draw a picture showing what it was like taking the test. You can draw a picture of anything that you would like; how you felt while you took the test, what you thought about while taking the test, or anything about the test. After you draw the picture, you will show it to me, and I will give you a sticker to put on your picture.

Stimulated recall interviews. After each child took a test, one of the researchers fluent in the child's L1 asked the child, in his or her native language, a series of questions about the test. The English interview questions are in the Appendix.

Procedure

We recruited in the Greater Lansing area and on the Michigan State University campus. We asked parents through flyers, email, and word-of-mouth to volunteer their children. Each child and at least one parent or guardian met with one of the researchers twice. Each session lasted about one hour. During the first session, the parent and child signed consent forms (which are in the supplemental file that can be obtained by emailing the authors). The researcher then had the child take the Bronze test, and asked the parent or guardian to fill out the background questionnaire for information on the child. The parent or guardian stayed in the room while the child took the test. After the child finished the

test, the researcher asked the child (in his or her native language) to draw a picture of how he or she felt when taking the test (see the directions above). After the child drew the picture, the researcher had the child pick out a sticker for the picture, and then asked the child a series of questions (Appendix) about his or her test-taking experience. During the test-taking, picture-drawing, and interview parts of the session, a second researcher videotaped the child.

During the second data collection session, which was one to three days after the first session, the child took the Silver reading test. After the Silver test, the researcher asked the child to draw a picture, as after the Bronze test. After the child picked out a sticker for his or her picture, the researcher asked the same interview questions (Appendix), but this time about the Silver test. A second researcher again videotaped the child as in the Bronze test session.

At the end of each session, the researcher had the child select a toy from the project's treasure chest. (Toys were valued at up to \$10.00 each. They included puzzles, board games, art kits, and science experiment kits.) After the second session, the researcher gave the parent or guardian a \$50.00 gift certificate to a major retail store. Thus, most children received two toys, and the parent or guardian received one gift certificate. Participant 7 did poorly on the Bronze test and also appeared rather stressed during testing. She picked a toy after her Bronze test. We did not have the child come in for the second test. The parent was mailed a \$25.00 gift certificate.

Analysis

We used three programs to analyze the data: Microsoft Excel 2010 to calculate item and test difficulty/facility levels and item/test discrimination levels; IBM SPSS (version 22) to calculate reliability and inferential statistics; and NVivo (version 10) to analyze the qualitative interview data and the children's drawings. We imported the interview videos and the drawings into NVivo. For the interview videos, we first transcribed them in NVivo (using NVivo's transcription fields), and if any of the interviews or parts of the interviews were in Chinese or Korean, we translated them into English.

To create a coding system for the interview data, we used an *inductive approach* (Thomas, 2006), which allowed the codes to emerge from the data. However, the interviews were not pure *stimulated recall* interviews as defined by Gass and Mackey (2000). We did have a stimulus; we allowed children to look at their test

booklets and/or their drawings during the interview. But, different from most stimulated recalls, which do not guide participants through a discussion, we guided the children through a list of standard questions (see Appendix), which we hoped would shape the interviews and help the children talk more about their test-taking experiences. Thus, the themes we created emerged from the data, but the data themselves were stimulated by revisiting the test booklet *and* by our questions.

We imported the drawings into NVivo as image files, and coded them on *tone* (adapting from Carless and Lam's 2014 work), that is, whether the drawing appeared to have a negative tone, a neutral tone, or a positive tone (examples are in Figure 1). We also coded whether for each child there appeared to be a change, as seen through the drawings, in difficulty between the two tests, and if so, we recorded the direction of that change (an indication that the Silver test was more difficult, no indication that one test was harder than the other, or an indication that the Bronze test was more difficult). We also took notes on the drawings, and cross-checked our notes with what the child actually said during the interview when asked, "What were you thinking when you drew this? What is this picture about?" All drawings

were coded by at least two researchers. We discussed discrepancies in coding (although there were only 3) until we reached a consensus.

To analyze the interview data, using NVivo, we first created 19 person *nodes* (NVivo's terminology for a coding category) that represented the 19 participants, and we classified each of the 19 person nodes with the following *attributes* (another term used by NVivo): age at time of testing, L1, family income level, length of English instruction in months, and length of residence (LoR), also in months. When a classification did not pertain to a person (such as *length of English instruction* for a native speaker), we used the term *not applicable* (a default category available in NVivo). We next coded the data that (a) was a child describing his or her drawing or (b) was a child comparing the Bronze test to the Silver test. We additionally coded all materials from a child as relating to either the Bronze test or the Silver test. We coded the data for preferences: If the child was describing something he or she liked, we coded it as "like," and if he or she was describing something he or she disliked, we coded it as "dislike." If the child was talking about a certain part of the test, we coded the data as referring to that specific test section (sections 1 through 5 for the

Table 2: Inherent coding categories and emergent themes we used to code the data

Code Type	Primary Code Level	Secondary Level	Third Level	Fourth Level
Inherent	a. Participant	I. - XIX. (1–19)	i. Age	7, 8, 9
			ii. L1	Chinese, English, Korean
			iii. Family income	0–25K through 175–200K
			iv. Length of English instruction	in months
			v. Length of residence	in months
	b. Drawing description			
	c. Test comparison			
	d. Test	I. Bronze		
		II. Silver		
	e. Preference	I. Like		
		II. Dislike		
	f. Test part	I. Pictures		
		II. Directions		
		III. Section	i. Bronze	Part 1, 2, 3, 4, 5
Emergent			ii. Silver	Part 1, 2, 3, 4, 4b, 5, 6
	g. Task unfamiliarity			
	h. Confusion			
	i. Counterfactual			

Bronze test; sections 1 through 6 for the Silver test, with the final summary question in Silver section 4 separated out as section 4b). We further distinguished if the child was talking about a test's directions or pictures. After we finished those larger, relatively inherent coding tasks, we inputted the following codes (or themes) into NVivo as thematic, emergent *nodes* that appeared to have risen out of the data: (a) a problem or discussion that pertained to the child's *task unfamiliarity*, (b) evidence of *confusion* about something on the test; or (c) a discussion of a *counter-factual item* (that is, the item states something that is not true; the child must recognize the fallacy and mark the statement as false). We coded for these three themes in particular because while watching the videos through the first time, we concurred as a group that these themes appeared to reoccur in diverse contexts, for different children, and at varying times. We wanted to see, by using queries and cross-tabulations with the inherent categories in NVivo, if there would be any patterns to these three emergent themes' occurrences, and if any subthemes to the larger three, emergent themes would appear. An outline of the coding is in Table 2.

Results and Discussion

Quantitative Results

Before answering the research questions, we first calculated descriptive and correlational statistics (in IBM SPSS, version 22) and performed classical test analyses (in Excel 2010) on the data to understand the data better. We first recorded all of the learners' individual item scores as right (1) or wrong (0) in an Excel spreadsheet. We tallied each learner's score on each test, calculated a composite score for each learner across the two tests, and followed common methods for calculating item facility and item discrimination (see Carr, 2011), using the native-speakers of English as the upper-level or *expert* group (the group expected to do well) and the nonnative speakers as the lower-level or *novice* group (the group expected to perform not as well as the expert group). We also recorded in the Excel spreadsheet the exact response given by each learner on each item. We summarize these data next.

In relation to research question one, we found that the Bronze test was indeed (as expected) easier than the Silver test. The Bronze's facility was 80% for the 19 learners who took it (facility climbs to 83% if Participant 7 is removed from the Bronze-test data). The Silver test's

overall facility (for the 18 who took it; Participant 7 did not take the Silver test) was 73%. Using the test scores from everyone except Participant 7, we performed a paired samples *t* test. To run the *t* test, first we derived the proportion correct for each test taker on each of the two tests. (The SD derived from the proportion correct Bronze scores by the 18 participants = 0.14; Silver SD calculated in the same way (minus Participant 7 and on the proportion correct scores) = 0.18). From the *t* test, we found the Silver test is significantly harder than the Bronze test, $t(17) = 3.511$, $p = 0.003$, Cohen's $d = 0.45$, effect-size $r = 0.22$. The Bronze test only minimally discriminated the nonnative speakers from the native speakers (overall test discrimination index = 0.12), and this could be because the Bronze test was rather easy and the range of scores rather narrow and toward top of the scale. The Silver test discriminated between the two groups better (discrimination = 0.23). The Bronze test had five (out of 25) items that had discriminating powers at 0.25 or above (items 5, 11, 14, 18, and 25). Meanwhile, the Silver test had 18 such items (out of 40) based on the same criteria (items 8, 12, 21–32, 34–36, 38). Looking at these statistics, one might suppose that the Silver test is a better test for this sample population of students because (a) it is more (appropriately) difficult for the entire population (less skewed data overall; fewer students at ceiling), and (b) its overall discriminating power is higher. But we must warn that the results should be interpreted with caution because the sample size is low.

Next, we looked at whether any of the learners' background or family characteristics correlated with test scores. We did this to explore the data more; we have to caution here that because our participant numbers are low, the results we found are to be interpreted with caution. Interestingly, we found (as shown in Table 3) that the learners' age ($\rho = 0.564$, $p = 0.012$) at the time of testing and grade in school ($\rho = 0.586$, $p = 0.008$) correlated strongly with their outcomes on the Bronze test, which one would expect (at least for the native speakers) because age mostly corresponds with grade in school ($\rho = 0.877$, $p = 0.000$). However, these correlations were weaker and non-significant in the case of the Silver test. In other words, the Silver test scores were associated with age and grade to a lesser degree ($\rho = 0.306$, $p = 0.216$; $\rho = 0.338$, $p = 0.171$, respectively), such that the correlations themselves were insignificant (not generalizable beyond the current data set). Correlations, as an inferential statistic, are highly susceptible to non-significance when the sample is small

Table 3: Correlations (Spearman's Rho) between student/family background and test score

Variables	Age at Test	Grade in School	Parent Education	Books Read (or read to) Per Day	Bronze Test Score	Silver Test Score
Grade in School	0.877**					
Parent Education	-0.180	0.104				
Books Read (or read to) per Day	-0.100	-0.090	-0.028			
Bronze Test Score	0.564*	0.586**	-0.192	0.172		
Silver Test Score	0.306	0.338	0.143	0.094	0.580*	
Bronze & Silver Test Scores Combined	0.427	0.472*	0.069	0.149	0.758**	0.973**

Notes: N = 19 for all correlations, except for those with "Silver Test Score:" 18 learners total took the Silver test. **Correlation is significant at the 0.01 level (2-tailed). *Correlation is significant at the 0.05 level (2-tailed).

and when the correlation coefficient is not large. In this data set, due to the small number of participants, a relationship between two variables *cannot* be detected as significant (generalizable) unless it is a medium-sized one (with a rho above, approximately 0.45). Thus, we believe that with more test takers, these correlations may have proven significant.

In relation to research question two, we found that almost all of the items on the two tests functioned as expected: native speakers did better on almost all of them than nonnative speakers did. Contrary to expectations, we found that three items (16, 17, & 22) on the Bronze test (out of 25 items total) and five items (2–5, 33) on the Silver test (out of 40 items total) were more difficult for the native speakers than for the learners of English. Later in this report, we discuss these items. Where we can, we use retrospection and interview data to give explanations as to why the items were inversely discriminating. But first we present how we coded the qualitative data, which included the interview data and the pictures that the test takers drew.

Qualitative Results

Findings based on the children's drawings

We asked each child to draw a picture after each test. We used this picture-drawing task (adapted from Carless 2012; Carless & Lam, 2014) because we wanted to better understand how the children felt while taking the test. We believed, as Carless and Lam did, that some children may be shy about answering questions posed by us (i.e., authority figures that the children did not know). Much work in child psychology has shown that

picture drawing is a familiar task for young children and one in which they feel at ease to express themselves (Carless & Lam, 2014). Indeed, many of the children in this study did appear to be able to express their feelings about the given tests through their drawings. Their drawings, we think, demonstrated their anxieties, their perceived successes, their perceptions of the test-taking environment (with drawings and comments on the desks, chairs, the proctors, their parents) and what they viewed as distractions in the room (e.g., their own hair, the windows to the outside, the treasure box with toys).

To recap the test-score data, we found the Bronze test was (as expected) easier than the Silver test (answering research question one). The children's drawings and the interview data appeared to corroborate these quantitative findings. We coded 19 drawings based on the Bronze test, and 18 based on the Silver. (See Figure 1 for samples; contact us via email for a PDF portfolio with all 37 drawings.) Out of 19 children, 12 (5 of the 7 native-English speakers and 7 of the 12 English-learners) drew pictures that appeared to show a positive overall tone, happiness, or test-taking ease during the Bronze test. (See Table 4.) With the Silver test, children appeared to experience more difficulties overall, with 11 of the 18 children indicating through their drawings that the Silver test was more difficult or stressful than the Bronze test.

The interview data was helpful in triangulating the picture-drawing data. We conducted 37 (19 Bronze and 18 Silver) interviews. During 33 of those interviews, the children discussed with the researcher his or her drawings, explaining what he or she drew and why. Thirteen out of the 18 children who took the Silver test

explicitly compared the Silver test to the Bronze test. And out of those 13 children, one (Participant 2, an English-speaking 7-year-old) stated that she thought “they are both kind of equal” in terms of difficulty. One other child (Participant 18, Korean, 9-years-old) was vague in her comparison; we did not code her interview response for directionality (“It took some time for me to take the Silver test because I had to think about the answers” (생각하느라 시간이 조금 더 걸렸어요.)), although her drawings did appear to more clearly indicate that the Silver test was more difficult. The remaining 11, however, explicitly stated that the Silver test was more difficult than the Bronze test. Indeed, we conclude that most of the children could recognize the tests’ different difficulty levels. Examples of what the children said are below:

Example 1 Participant 3, L1-English, age 8

-
- Researcher: Okay. Let’s talk about the [Silver] test. So, how was the test?
- Child: A lot harder.
- Researcher: Hard? What made you think it’s hard?
- Child: Hmmm, because, hmmm, there were more questions than the last time. Last time there were like twenty five, this time there were like forty.
- Researcher: More questions?
- Child: [Nods her head yes.]
-

Example 2 Test taker 15, L1-Korean, age 8

-
- Child: 전에 것보다는 어려웠어요. 그래도 아직도 평소에 하는 test는 그림 자체가 없거든요.
- It was harder than the last one. But it was not bad because usually the tests that I took [back home in Korea] do not have any pictures.
-

Example 3 Participant 1, L1-Chinese, age 9

-
- Researcher: 那这个测试怎么样呢
What do you think of this test?
- Child: 这个比上次要难。有些地方我不会做，有的单词我不认识。
- This one is more difficult than the last one. There are some questions that I do not know how to answer, and there are some words that I do not know.
-

The data summarized in Table 4 may help us better understand to whom the two tests (Bronze versus Silver) should be administered. In this data set, it appears to us that Participants 7, 14, and 12 should *only* be given the Bronze test (and not the Silver test). Our rationale for that recommendation is based on the test takers’ proportions correct (out of 25 on the Bronze and 40 on the Silver test). For those three test takers, the Bronze test was most likely more developmentally appropriate and, perhaps, a better measures of those test takers’ levels of proficiency. When tests are too difficult for children, the children may feel inadequate, humiliated, stressed, or they may even question their own self-worth (Menken, 2008; Schmidt, 2000; Winke, 2011). Indeed, we stopped Participant 7 from taking the Silver test because she broke down during the Bronze test (because the Bronze test was too hard for her). The qualitative interview data revealed that Participant 16 wanted to stop taking the Silver test because it was too hard for her (she got a 22 out of 40 on the Silver test). When asked by the researcher “How did you feel when taking the test?” (기분이 어땠어요?), she replied, “Not good. I wanted to stop taking the test, but I didn’t” (안 좋았어요. 끝내고 싶었는데 안 끝냈어요.).

Findings based on the interview data

After we coded the interview data, we ran matrix queries in NVivo on the qualitative data to understand who talked about the various coding categories (by age of test taker, by L1 background, etc.), and in relation to which test (Bronze or Silver) and, further, in relation to which sections and items. One of the most fruitful queries we ran was a simple count of the children’s likes and dislikes by test and by test section. We were able to do this because we asked the children what they

Table 4: Codes we gave to the children's drawings and the children's test scores

ID #	L1	LoR (Months) in USA	Bronze drawing code	Silver drawing code	Did the drawings appear to indicate the Silver test was harder?	Did the child say (in interview) the Silver test was harder?	Bronze Test Score*	Silver Test Score*	Total Score
5	English	NA	3	1	Yes, Silver harder	Yes	24	36	60
11	Chinese	5	2	1	Yes, Silver harder	Yes	24	36	60
19	English	NA	3	1	Yes, Silver harder	DC	22	38	60
6	English	NA	2	1	Yes, Silver harder	Yes	25	34	59
1	English	NA	3	3	No, no difference	No, equal	22	35	57
2	English	NA	2	2	Yes, Silver harder	Yes	21	35	56
3	English	NA	3	1	Yes, Silver harder	Yes	20	35	55
18	Korean	3	3	2	Yes, Silver harder	Yes	23	32	55
17	Korean	3	3	1	Yes, Silver harder	Yes	23	30	53
4	English	NA	3	2	Yes, Silver harder	Yes	20	31	51
9	Chinese	7	3	2	Yes, Silver harder	DC	22	29	51
10	Chinese	5	2	1	Yes, Silver harder	Yes	23	27	50
8	Chinese	6	3	3	No, no difference	DC	20	29	49
15	Korean	8	2	2	No, no difference	Yes	23	25	48
13	Korean	10	3	3	No, no difference	Yes	18	23	41
16	Korean	3	1	1	No, no difference	Yes	15	22	37
12	Chinese	2	3	3	No, no difference	DC	12	17	29
14	Korean	10	1	3	No, Bronze harder	DC	17	12	29
7	Chinese	1	3	NA	NA	NA	7	NA	7

Notes: For drawing code, 1 = Negative tone or affect, 2 = Neutral tone, 3 = Positive tone; LoR = Length of residency, NA = not applicable; DC = Child didn't directly compare the two tests during the interview. *The Bronze test had a total of 25 items (for 25 points possible), while the Silver test had 40 items (40 points possible).

liked and disliked on the tests. Accordingly, most of the children discussed the tests and their sections in those terms. Looking at these data (Table 5), we see that the children (as a group) appeared to have the most negative opinions about Bronze test section 3 (9 dislikes, 0 likes), Bronze test section 1 (9 dislikes, 5 likes), and Silver test section 5 (10 dislikes, 4 likes). At the same time, some children both liked and disliked these individual test sections (except for Bronze test section 3). Examples from the children's transcripts reveal how this is possible (e.g., Participant 5 discussing Silver Part 5: "It is probably the hardest one and the longest one, but it tells a pretty good story, hmmm, about the kids and how the kid is scared of sharks;" Participant 10, discussing the same section: "I like the story, not the question" (我是最喜欢它的情节, 又不是这个题。)).

Below, we discuss the qualitative data primarily in relation to these three sections on the test that appeared to be the most divisive (Bronze section 3, Bronze section 1, and Silver section 5). In addition, we discuss three larger themes that emerged: (1) task unfamiliarity, (2) problems with counterfactuals, and (3) confusion. In doing so, we answer research question three: We found children did sometimes lose attention on the tests.

1. Task unfamiliarity. We believe the children's problems with Bronze test section 3 are complex, but overall the difficulties (and dislikes) are related to the children's unfamiliarity with the task. Bronze test section 3 requires the children to look at five simple pictures of clothing, shoes, or eye glasses and look at letter combinations next to the pictures. The children unscramble the letters to spell the word next to the picture. Each word blank includes the exact number of

Table 5: Count of which test sections children explicitly stated they liked and disliked during the interviews

Test	Section	Dislike		Like	
		N	ID#s	N	ID#s
Bronze	1	9	3, 5, 6, 9, 10, 12, 13, 14, 15	5	6, 9, 13, 15, 19
	2	3	2, 13, 19	5	1, 4, 11, 13, 18
	3	9	1, 2, 3, 9, 10, 11, 16, 18, 19	0	
	4	7	3, 4, 13, 14, 15, 16, 17	4	5, 14, 17, 18
	5	4	4, 14, 15, 17	6	3, 5, 8, 10, 12, 16
Silver	1	1	10	4	4, 16, 17, 18
	2	5	1, 4, 5, 9, 19	5	1, 4, 6, 12, 15
	3	2	4, 12	2	11, 19
	4	4	4, 12, 15, 16	1	18
	4b	1	1	0	
	5	10	2, 3, 4, 5, 6, 10, 11, 12, 17, 19	4	5, 8, 9, 10
	6	4	4, 5, 14, 15	3	1, 2, 14

Note: Shaded areas are those in which 9 or more children had a specific opinion (in one direction) about the test section.

spaces (one space per letter) required to spell the word. For example, next to a picture of a pair of brown shoes is a blank with five spaces (_ _ _ _) and the letter combination “esohs” that needs to be unscrambled. The children should write “shoes” on the blank with five spaces (one letter per space).

Comments from the children indicate they were unfamiliar with this type of exercise. For example, the children noted that they had trouble with (a) identifying the exact word for the picture shown, (b) cognitively processing the scrambled letters as scrambled letters (Participant 2 indicated she tried to read the scrambled letters as actual words), and (c) spelling the words, even with the letter clues (ultimately, some children did not use the scrambled letters to spell the words). We believe the test designers intended the scrambled letters to serve as clues, but instead, the comments demonstrate that for many children, the task was unfamiliar and perhaps overly taxing. In other words, because they had not experienced this task before in their schooling, they had to spend much time figuring out what to do, and this may have caused stress (as seen in some of the children’s picture-drawings). But ultimately, most children were able to complete this section, even though they did not like it.

Example 4 Participant 2, L1-English, age 7

-
- Child: I think the most difficult part is this part (pointing to Part 3) when I was spelling “jacket” [she spelled it “jackeat” and wrote in an extra space for the “t”] because there wasn’t enough space. [Turns the page and points to the page.] This mixed up stuff (laughs).
- Researcher: Why?
- Child: Because I thought this was going to be pajamas (laughs).
- Researcher: Anything else?
- Child: Ah I don’t wear glasses. I don’t even wear jackets. I do not wear sneakers (laughs).
- Researcher: So you are not familiar with these things?
- Child: No. [Pointing at Part 3.] Like

Example 8

Participant 6, L1-English, age 9

-
- Child: I got a bit confused like [pointing to Part 4, Bronze, pointing to word bank] this page I didn't see that [waves over the word bank]. I guess I saw through the instructions because, like, 'cause, like, I didn't really get it. I didn't know what this [points to word bank] is for. So, I was just like got confused and so I just like, did it. Then, I found out that was kind of like, word bank, but you were not supposed to use all the words because pianos are not related to horses. (laughs)
- Researcher: Were the instructions on the test clear?
- Child: Well, every single one [instructions] was clear except for this one [points to Part 4].
- Researcher: Can you explain it to me?
- Child: I think I kind of rushed through it, all this [points at the directions], 'cause like, "choose the words from the box?" Except this is really bad. I don't think I saw it [covers up the word bank with her hand]. So this was like, what? So I was like, where? What box? Then, I looked at THIS [points to the picture of a horse that is a box at the top of the page] and so I was like (laughs) maybe . . .
- Researcher: So you thought this [the horse picture at the top of the page] was a box.
- Child: Yeah, so I got like really confused.
-

Example 9

Participant 14, L1-Korean, age 7

-
- Researcher: 이 시험 중에 어떤 게 제일 좋았던 것 같아?
Which part was your favorite?
- Child: 왜냐면 이거 학교에서 했거든요. 이거는 몰랐어요.
I liked part 4 because I do this type of activity at school. But I don't know what this [word bank] is for.
- Researcher: 이거는 왜 있는걸까?
Why do you think this is? [word bank]
- Child: 몰라요.
I don't know.
- Researcher: 시험이랑 관련있을까?
Do you think it's related to the test question?
- Child: 어떻게 하라는 건지 써져 있지를 않잖아요.
I don't know because there is no information about it. There is no information about how to use it.
-

Example 10

Participant 19, L1 Korean, age 9

-
- Child: 저는 이게 쉬운데 이게 없었다면 (word bank) 아마 조금 더 어려웠을 거예요.
It would have been a bit more difficult if there were no word bank.
- Researcher: 처음엔 이게 없는 줄 알았어?
So you didn't know the word banks were given at first?
- Child: 네. 처음에 없는 줄 알고 첫 문장을 읽었을 때 무엇에 관해 말하는지 몰랐는데 이걸 보고 (word bank) 알았어요.
Yes. I wasn't aware of that, so I was
-

confused when I read the first sentence of the reading passage in part 4. But I was able to fill in the blanks after seeing the word bank.

왜 여기에서는 horse 에 대해서 말하고 있는데 여기에서는 (word bank) 왜 hippo 가 나오는지?

The reading in Part 4 is about a horse, but I don't understand why the word "hippo" is given in the word bank. It is not related to the reading passage.

여기 있는 단어들은 여기 있는 것들을 (reading passage) 위한 보기 단어들이잖아요. 그런데 여기에서 갑자기 hippo 나 piano 가 나온다는 것은 조금 그랬어요.

The words given in this box are possible options for the blanks in the reading passage, but I don't understand why they are giving words like "hippo" or "piano."

2. Problems with counterfactuals. In examples 8 and 10 above, the children indicated they were confused when vocabulary words unrelated to the reading passage were offered as possible (but incorrect) options in a word bank. These examples may demonstrate that in the children's school-based, L1-reading tasks, they are rarely presented with counterfactual information or incongruities. By third grade (age 8), children are reading to learn, not learning to read. And even when learning to read (in the younger grades), teachers most likely rarely have them read counterfactual information to test their reading comprehension. From the very beginning of child preliterate development (e.g., babies' basic picture books), children see large, colorful pictures with text that explains exactly what is seen in the pictures. Children's books normally do not present text that is incongruous with what is being shown in the pictures.

We found a number of examples of test takers being confused by counterfactual evidence when we analyzed the children's problems with Bronze section 1. In that section, children were to mark whether they thought the statement in relation to the picture was true (check) or false (X). In the first example sentence, a picture of a flower is shown, and the sentence next to it reads, "This is a flower." In the box next to the sentence is a check-mark indicating "yes, this is a flower." The second example is a counterfactual, with a picture of a cow and

the statement "This is a goat." An X appears next to that example sentence.

Five pictures and statements are given on this part of the test, and this section is worth five points. Two out of the five are counterfactuals (question 1, "This is a lizard," picture of a spider is shown; question 4, "This is a television," picture of a cell phone is shown). These counterfactual statements apparently caused much amusement and confusion, as indicated in the following examples.

Example 11 Participant 2, L1-English, age 7

Researcher: Which part was the [sic.] especially difficult?

Child: [Points to Part 1, Bronze, page 1, picture of spider] On this spot, I thought this spider was a lizard! Because it says, "This is a lizard!" (laughs) No!

Researcher: You thought that was difficult?

Child: No, I did not. But . . . this is NOT a lizard.

Example 12 Participant 5, L1-English, age 9

Child: These were the most obvious [pointing to Bronze Part 1], like [laughing], why would you be like, getting to test, like, [pause], this does NOT look like a goat. That is not a goat. And this is obviously not a lizard. It's a spider. It's an arachnid, not a reptile. It has like eight legs instead of four. A lizard has a tail. There are some things that are the same between spiders and lizards. Like, they could both be poisonous. But they didn't ask what it was. And I don't know [points to third picture of a phone on page 3] what that was. That's not a television.

Researcher: Were there any unclear parts in the pictures?

counterfactuals in section 5 of the Silver test where the problem was more subtle with text/picture incongruences. In other words, the children appeared to have expected that the pictures would align exactly with the text (as stated by Participant 5 in Example 12 above), but some of the pictures were missing certain objects or did not exactly show people's expected expressions. These slight misrepresentations in the pictures were bothersome to some of the children.

Participant 4, L1-English, age 8

Researcher: Let's talk about the pictures then.
Were any of the pictures unclear?

Child: [looks through test booklet, points to picture on page 11, in Part 5, Silver test] This one.

Researcher: Hm. Why was that unclear for you?

Child: Well, cause, Paul doesn't want the ice cream, but he's still smiling!

Researcher: Okay, he's still smiling. Okay.

Child: And the mom, she's looking, but she's still in the water.

Child: 좀 웃긴 것은 있었어요. 이런거
이런거 엑스 돼 있는 거는 너무
웃겼어요. 왜 이게 lizard 인지
모르겠어요.

This part was funny (Part 1, the “spider”
picture). I thought this example of
putting an “x” next to this sentence was
funny. I don’t know why this is a lizard.

The same child also pointed out that in section 3 of the Silver test, the directions stated that “Peter is talking to his friend Jane,” but clearly, in the picture just above the directions, it is Jane who is talking to Peter. While these may be seen as slight quibbles about the test, they point to a potentially larger issue. Standardized tests probably should not require very young children to identify false statements, even though such items have been identified as acceptable if they are carefully thought out and pretested (McKay, 2006, pp. 240–241). Indeed, the *Common Core State Standards Initiative* (2010) has indicated that students should be able to “identify false statements and fallacious reasoning” beginning in Grade 9, when children are approximately 14 years old (p. 40). Nor, according to our data interpretations, should very young children be presented with pictures that do not align well with the associated text. These are not situations very young children normally encounter in real life.

Page 16

their inability to understand the directions, which led to confusion. In fact, *dislikes* were often intertwined or overlapping with *confusion* in our data set. This inability to comprehend is likely related to task unfamiliarity. As some of the directions were simple, and yet confusing to the children, we feel it is important to point out the children's problems. The confusion does make sense when explained by the children. The directions that they appeared to have the most trouble understanding (based on a matrix query of *dislikes*, *directions*, and the *test sections*) were these: Bronze, Part 4 (eight children had trouble); Silver, Part 5 (four children had trouble); and Silver, Part 6 (six children had trouble).

Bronze Part 4 directions. In Bronze Part 4, the first word bank appeared which we believe was the source of much confusion. In Example 8 above, Participant 6 discussed her struggle understanding the directions: "Read this. Choose a word from the box. Write the correct word next to numbers 1-5. There is one example." She did not understand *to what* the word "box" in the directions referred, and she first thought the box at the top of the page (with a picture of a horse in it) might be the referent, but that, she said, made her even more confused. Similarly, Participant 15 indicated she did not understand Bronze Part 4's directions. She said, when asked if any of the instructions were unclear, "Part 4's instruction. Here it says "one to five," but there are actually 10 examples down here [in the word bank]." (이거요. 왜냐면요. 이게 one to five 라고 하는데 example 이 10개가 있는데 one to five 라고 하니까.). Others were confused with the word bank because it had extra pictures and words (distractors) which would not be used (see Examples 8 and 10 above, in which the test takers described their confusion over the existence of the distractors).

Silver Part 5 directions. The directions to Silver, Part 5, stated (in full): "Look at the pictures and read the story. Write some words to complete the sentences about the story. You can use 1, 2, or 3 words." A few of the native and nonnative speakers indicated they did not understand what "1, 2, or 3 words" meant. This may be because they had never had to count the number of words they wrote in response to reading-comprehension questions before, a type of task unfamiliarity. Thus, to them, the sentence "You can use 1, 2, or 3 words," may have appeared as gibberish (they read it, but they could not process it).

Example 16

Participant 10, L1-Chinese, age 8:

Child: 我在学校考试的时候，都有例子。但是这个考试有的有例子，有的没有。而且有点不习惯。它只告诉我们填一二或者三 【指第五部分】。不是很清楚。

All the other tests I took in the school, they all had example. But, this test, some of questions had examples, some did not. I think they all should have examples. It only told us to fill in "1, 2, 3" (pointing Part 5). It is not clear.

Example 17

Participant 4, L1-English, age 8

Researcher: Let's talk about the test instructions. Were they clear to you?

Child: [points to instructions on page 10, for Silver Part 5] I didn't really understand "You can use 1, 2, or 3 words."

In Silver, Part 6, there was a word bank, but the directions ("Read the text. Choose the right words and write them on the lines.") did not explicitly allude to the word bank or the need to use words from it, thus most of the confusion was by children who discovered the word bank on their own and then may have re-read the directions, trying to figure out how to use the word bank.

Example 18

Participant 15, L1-Korean, age 8

Researcher: 제일 어려웠던 부분?
Which part was the most difficult for you?

Child: 여기요. Confusing 했어요.
Part 6. It was confusing.

Researcher: 어떤 부분이요?
Can you explain about that?

- Child: Direction 이 아무것도
없어가지고.
There were no directions on how to
answer the questions.
- Researcher: 다 풀었는데 이게 나와서 이걸
읽지하고 다시 풀었죠?
Did you read the instructions in this part?
- Child: [reads directions again]
Text 가 뭐예요.
What does “text” mean in Korean?
- Researcher: Text 글자 써 있는 거
Text means this, like the reading passages
and sentences.
- Child: 어떻게 choose the right word 에요?
옆에서 보라는 말이 없는데?
But how do you choose the right word
when there is no instruction to look at
the right part of the page?

These data demonstrate that even after pilot-testing (the YLTE tests were piloted extensively by Cambridge English before they became operational) children may still struggle with test directions, especially if they do not prepare specifically for the test, as in this study. During this study, when we acted as test proctors, we only answered questions about test directions. We did not check to see if children understood the directions. Mainly, we let children read the directions themselves. Indeed, the YLTE directions to proctors document states that once the test is started, proctors cannot answer questions, even about the directions. Proctors, we assume, are *not* normally instructed to *ask* if children understand. They do not normally *check* for comprehension of the directions. Our data show that allowing teachers or proctors to explain test directions to young children is extremely important, but perhaps more important than that would be to ask teachers or proctors to ask each child if clarification is needed. This may involve a rethinking of the role of proctors in standardized, high-stakes, child language assessment. In Example 19, Participant 19 explains how she was confused even after she had read the directions.

Example 19 Participant 19, L1-English, age 7

- Researcher: Can you show me the part you didn't like? What did you not like about it?
- Child: Well, [points to part 5 in the Silver test] I couldn't really understand it. I couldn't really understand what I was supposed to do, even though they explained it.
- Researcher: Can you show me which part was hard, or any more information you can give me? Because if other children are going to take this test, what might be difficult for them, if it was difficult for you?
- Child: Well, if they're older children, I don't know if they will struggle, but if they are younger, then I think you should do some talking to them in words, about this [pointing to the directions].

These data brought up new questions for us. Should children spend time preparing for the test so that they become familiar with the test format and directions? In our study, should we have given practice tests to the children beforehand to help ward off confusion about the directions? The data appear to suggest that children should take at least one practice test. On the CaMLA website, parents and teachers are instructed that “CaMLA does not prescribe or endorse any specific course of study to be taken in preparation for the YLTE . . . The best preparation is through general study and use of English.” However, parents and teachers are instructed that they “may find it useful to consult the complete sample tests available on our website.” If children do become better prepared through practice testing (if practice testing helps reduce construct-irrelevant score variance or even confusion during testing), then we believe that CaMLA may want to more strongly suggest that children take at least one sample test prior to taking a real YLTE test. Doing so may help the children become *testwise*.

Testwiseness has long been investigated in language assessment research. It is described as being able to apply appropriate and effective test-taking strategies that relate

directly to the test format (Sarnaki, 1979). Testwiseness may help children maximize their observed test score (Rogers & Yang, 1996), even though it is considered independent of the test takers' knowledge of the subject matter being tested (Millman, Bishop, & Ebel, 1965). When accrued through test preparation, testwiseness may help students implement metacognitive strategies appropriate for and relevant to their language proficiency levels and in relation to the test items (Cohen, 2007). Additionally, testwiseness through test preparation may help lower test-taking anxiety through basic test-format familiarization (Winke & Lim, 2014). Problems in interpreting score reports may occur if testwise students' scores are not differentiated from non-testwise students' scores. Research is needed on how test preparation affects children's test-taking experiences and test scores.

Interview data in relation to the inversely discriminating items

Quantitative data revealed that three items (16, 17, & 22) on the Bronze test and five items (2–5, 33) on the Silver test were inversely discriminating. Here, we discuss those items in relation to the students' interview responses.

Bronze items 16 & 17: A potential problem with option plausibility and background knowledge. Figure 2 is an example of items 16 and 17 from the Bronze test. These two items (items 1 and 2 in Figure 2) proved to be difficult for two of the native speakers because (as our qualitative interview data showed) they neither understood (task unfamiliarity) nor used the word bank. For example, in Figure 2, Participant 4, a native speaker of English, used “mane”: not the word in the key, “hair.” In essence, she was penalized for not following the directions, but one might argue that the answer she provided is actually better. It could be surmised that learners of English unfamiliar with word banks and/or with equally sophisticated vocabulary related to horses might also get this answer wrong by filling in “mane” as Participant 4 did, or by trying to fill in the word “mane” and misspelling it, as Participant 3 (also a native speaker) did. The same problem occurred with item 17 (number 2 in Figure 2). Instead of writing in “house” from the key, two native speakers, Participants 3 and 4, wrote in their own responses (Participant 3 wrote in “forest;” Participant 4 wrote in “cage”), which did not appear in the word bank. These data demonstrate that *reading* the instructions, for a child, does not equal *understanding* the instructions.

Bronze item 22: A potential problem with age-appropriate misspelling. Figure 3 provides an example of a free-response item which assesses the integrated skills of reading, scene perception, and spelling. Two native speakers (Participants 1 and 3) and one nonnative speaker (Participant 12) spelled the response “girl” wrong, inverting the “i” and “r” and producing “gril.” The misspelling meant that the three test takers got this item wrong. It may be that this type of spelling mistake (an /r/ before the vowel when it should come after) is common for young, native-English-speaking children because children often spell inaccurately, but in phonetically plausible ways (Bourassa & Treiman, 2001). And the younger the child, the more common it is for him or her to misspell in certain ways (Treiman & Cassar, 1997). Indeed, commonly observed misspelling patterns match children's natural process of acquiring phonological awareness (Treiman & Kessler, 2014). Apropos the error noted above (spelling the word *girl* as “gril”), Bourassa and Treiman (2001) noted that native-English speaking children when learning to read and write in English “treat nasals and liquids as qualities of the vowel that precedes them rather than phonemes in their own right” (p. 173). Additionally, Bourassa and Treiman indicated children have problems with the interior consonants of initial clusters, which is what appears with spellings like “gril.” In particular, Bourassa and Treiman indicated that the liquid /r/, like /l/, can be troublesome in this position.

To spell a word such as *far*, children attempt to divide the spoken word into individual sounds or phonemes and to represent each phoneme with a letter. However, the /ar/ sequence in this word is difficult to segment. As argued earlier, children tend to group vowels and following *rs*, treating them as a single unit (p. 175).

Mapping their argument onto the misspelling “gril,” we suggest that the children may be segmenting the word “girl” into two individual sounds: “gr” and “il.” But the question is, if it is natural for young children to misspell certain words in certain ways, and if this is part of L1-English children's normal process of learning orthographic patterns and morphological relations in English (Treiman & Cassar, 1997; Treiman & Kessler, 2014), can language testers treat the same misspellings by like-aged, nonnative speakers of English as evidence of English-language-learning deficits (or *low proficiency* in the language)? And if younger children are more apt to misspell a certain word, should an item requiring the

correct spelling of that word be weighted or scored the same for all child age groups? We suspect the responses to these questions may be *no*. Most certainly, we now are convinced that applied linguists involved in designing L2-English spelling items (to assess the English-language writing ability of young learners of English) must research and review the vast literature on children's L1-English-spelling development (i.e., Critten, Pine, & Messer, 2013; Nunes, Bryant, & Bindman, 1997; Treiman & Kessler, 2014). Language testers must do this to ensure that child test takers are given fair and age-appropriate spelling-test items. Items that are tricky for the test takers due to the test takers' level of cognitive development/age or that are above their expected levels of phonological awareness based on their L1 and L2 backgrounds should be avoided.

Silver items 2–5: A potential problem with focus on form versus meaning. Figure 4 contains an example of items 2 through 5 from the Silver test. The answer key indicates that the answers are correct only if the article, if needed, is included in the response. This resulted in native speaking Participants 2, 5, and 6 getting items 2 through 5 incorrect, even though they selected the correct nouns (meaning) for the items. Similarly, nonnative speaking Participants 13 and 14 got a few of these incorrect for not transferring the article. We asked one of the native speakers why she did not transfer the articles with the correct words (Participant 19), and she stated that she had never had to do that before. In other words, she didn't see the point in doing that, or why it would be important. Another potential reason that they may not have seen a point in including the article is that the answers are not part of a sentence; rather, they are instead just a blank after the question, so they may not have thought that there was a need to respond with anything other than the noun. Research suggests that children are more correct with article usage when they must use the articles in context (Zdorenko & Paradis, 2008).

Correct article use is a developmental process in children (Zdorenko & Paradis, 2012), with the development depending in part on the child's L1 (one that has articles, or one that does not) (Zdorenko & Paradis, 2008). Thus, it may be that it is unfair to test article usage in young children, as the children's extreme variation in article use may have more to do with their general age and L1 background than with their overall ability to comprehend and communicate in English. We did find it curious that the nonnative speakers tended to transfer the article. We believe this may stem from their

more formal instruction on English as a second language in school. They may have a greater focus on form (which may show up as greater attempts at form accuracy when tested), while native speakers of English may have a greater focus on meaning.

Silver item 33: A potential problem with age-appropriate over-generalization of grammatical rules. Figure 5 contains an example of item 33 from the Silver test. This was a free-response item, with the response to be based on the reading presented above. Two native speakers and three nonnative speakers left the plural marker "s" off the word "dolphin." We speculate that because this is a marine animal, some children might have assumed that "dolphin" without the plural marker "s" is an acceptable way to pluralize the noun (as with the words "fish" and "mackerel"), which may be an example of a common, age-related overgeneralization of a grammatical rule, something natural to the language acquisition process (Berko Gleason, 2004). Indeed, Pinker (2011) noted that it is common for children to overgeneralize almost any type of grammatical rule, but moreover, it is difficult to understand *why* children over-apply specific rules. He wrote, "overgeneralization errors are a symptom of the open-ended productivity of language, which children indulge in as soon as they begin to put words together" (Pinker, 2011, p. 190). Another explanation, however, is that some of the children copied the closest occurrence of the word "dolphin" from the text, which appears in the second-to-last line of the text, and which was spelled without the plural marker at that point in the text. Eye-tracking might reveal the true nature of this mistake. For example, if we had used an eye tracker, we could have seen whether children looked at "dolphin" in the second-to-last line of the text directly before writing the word "dolphin" (without a plural "s") on the line. The larger question is, if it is common for children to overgeneralize grammatical rules, how can language testers ensure that natural overgeneralization is not counted against the child in a test of English grammar? It is a complex question, but it is something that can be monitored by having grammatical items pretested on a sample of same-aged, native-speaking children.

Conclusion

In this study, we investigated the validity of two specific tests of reading for young English-language learners, the YLTE Bronze and Silver tests. All tests have measurement error, especially child tests (Biggar, 2005).

This is because children are non-ideal test takers (McKay, 2006; Menken, 2008; Pukett & Black, 2000). They may not understand why they are being tested, who will score their tests (one child in our study noted she thought her own teacher would score her test), or what the test outcomes will mean for them. They have little or no *testwiseness* (Carter, 1986; Millman, Bishop, & Ebel, 1965; Rogers & Harley, 1999; Sarnaki, 1979), and part of this might be because, as we showed in this study, young children can be novice test takers, unfamiliar with common test formats, test directions, and test tasks.

We followed Messick's (1989) suggestion to do more than just look at test data (to find concurrent or predictive validity). We also wanted to do more than seek experts' judgements about the tests to have evidence concerning the tests' content and construct validity. We wanted to look at the tests' validity from a different angle. That is, we wanted to learn more about how test takers responded to the YLTE test tasks, and we wanted to investigate their test processes. We tested protocol methods suggested by Green (2014), Weir (2014), and Carless and Lam (2014).

The first research question was whether the Bronze test was less difficult than the Silver test, as is intended. All data (test scores, interview data, and drawings) indicate that the Bronze test is indeed easier than the Silver test. A related question that we attempted to ask was *for whom* (within our sample) each test was best suited. We wondered if we would see a clear cut relationship in terms of demographics. Unfortunately our data did not point to any clear associations, except that students' ages and grades correlated with their scores on the Bronze test. (Older and higher-grade students did better on the Bronze test, and if we had had more participants, most likely we would have found significant results indicating that age and grade correlated positively with Silver test scores as well.) We did find that the Bronze test was too difficult for one test taker (Participant 7), and the Silver test was most likely too difficult for at least two additional test takers (Participants 12 and 14) who received scores on the tests that were below 50%. (Using that cut off, one could argue that the Bronze test was also too difficult for Participant 12.) It might be somewhat arbitrary to decide that a score below 50% indicates that a test was too difficult. Clearly, a 50% threshold would need to be determined on a test-by-test basis and in conjunction with information on the tests' uses. CaMLA provides direction in regards to when a student should move on to the next level: When a child performs well on a test

(as indicated through a scaled score report that uses a child-friendly medal system), he or she "should be ready to start preparing for the next CaMLA YLTE exam" (CaMLA, n.d.). Using raw test scores combined with qualitative interview data, we surmised that Participants 7, 12, and 14 should not have gone on to take the Silver test. But how could Participant 7 have been prevented from taking the Bronze test? The CaMLA website notes that for the Bronze reading test, the level of reading that Bronze test takers can do is "recognize the letters of the alphabet." Participant 7 could do that. She could also "write the letters of the alphabet and spell her name and simple words," but she performed poorly and appeared to be stressed during Bronze testing. We think the descriptors of what students at the Bronze level can do may need to be revisited. Additionally, we believe that if there is enough demand, there might be room for a test level below Bronze. Or, there could be a short screener test to see if children are ready for the Bronze test. Such a pretest or screener might help prevent children with too low of an English-language proficiency level from taking the full Bronze exam.

The second research question was whether native speakers would perform better on the tests than the nonnative speakers. We found on almost all items, the native speakers performed better than the nonnative speakers did. On the few items where the nonnative speakers performed better, we found reasons why this happened, and we also found possible solutions. The next step may be revising the test, amending the answer key, or changing proctoring protocols. McNamara (2000) explained that test validity arguments are like arguments presented in a court of law. However, unlike in a court of law, language-test validity cases are never closed. More arguments can always be made (Green, 2014). There is always a way to improve upon the quality of a test, and each validity argument can show how a test can be improved.

The third research question, which stemmed from prior research (McKay, 2006; Menken, 2008; Pukett & Black, 2000; Winke, 2011), was whether children lose attention during the tests. We found that overall, children did pay attention throughout the whole test. They mostly tried their best, which is not surprising given their parents were in the room. But they sometimes got distracted, as children their ages do. After all, children are cognitively limited in terms of what they can process and what they can do (Pukett & Black, 2000), and the testing-taking situation was stressful or difficult for some of the children some of the time.

Research has suggested that children at ages 7 to 9 are still learning to think logically, consider problems from different sides, and make inferences (Cook & Cook, 2009). This may explain why at least two children (Participants 1 and 11) had problems with the task to “Choose the best name for the story” (Silver test, Part 4b): neither child could choose the best name; Participant 1 wrote in the test booklet her own answer that was different from those given, and Participant 11 said he was confused because he thought all options were good. As explained by Cook and Cook:

By age 7 most children are capable of using logical thought structures . . . However, there is still one major limitation in their thinking: Their use of mental operations is still closely tied to *concrete* materials, contexts, and situations. In other words, if children have not had direct experience with the context or situation, or if the material is not tangible, they are not successful in using their mental operations. (p. 166)

Thus, we found some ways in which the test could be modified to prevent confusion and to make the test a little more valid for the test-taking population. But moreover, we speculate more good could be done by changing proctoring protocols for high-stakes or standardized exams involving young children. Normally proctors distribute test materials, supervise the testing room, supply materials, guard against cheating, and maintain seating charts. They collect materials afterwards, dismiss students, and might be involved in grading. CaMLA informed us that YLTE proctors can assist students with directions and answer general questions, but it is not clear if YLTE proctors (or protocols of any child language tests) can proactively ask students one-on-one if they understood. What would happen if proctors acted even more like teachers? What if they could circulate around the room verbally checking to make sure students understand the directions and tasks? Because children normally have their teachers checking in on them when they do tasks in the classroom, in testing situations, should test proctors do the same? This may need to be done if our results are indicative of other child language-test-taking situations. Even after students have read the directions and understood every individual word in the directions, they sometimes do not *process* the directions correctly, or they understand incorrectly, rendering wrong answers when they had the ability to get the answer(s) right.

We have two main limitations that we would like to discuss. First, the picture-drawing task, while extremely interesting in some cases, did not always work. Some children drew what they were thinking at the time of the drawing, and not what they thought while taking the test, which is not surprising because even adults sometimes have difficulties remaining retrospective during stimulated recall sessions (Gass & Mackey, 2000). For example, Participant 2 drew what she had last seen in the test booklet (three cats), something she continued to think about after having taken the test, not what she thought during the test or what she felt like during the test. Nonetheless, many of the drawings were a good triangulation of the interview data, and the picture-drawing task had (we believe) psychological benefits. The task allowed the children to relax after having taken a test. Nonetheless, researchers must ask students why they drew what they did, or ask students to write a caption for the picture as done by Carless and Lam (2014).

The second limitation is this: We believe we could have included other levels of data analysis. We could have used NVivo to code the children’s test booklets. That is, the children’s test pages could have been scanned in as image files in NVivo, and erroneous responses could have been coded for possible reasons (other than English ability) for the wrong responses. Additionally, we believe that eye-tracking data could be particularly important to better understand the children’s thought processes while they take child language tests. We hope to see such data collection with children in future validation work.

References

- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Berko Gleason, J. (2004). The child’s learning of English morphology. In B. C. Lust & C. Foley (Eds.), *First language acquisition: The essential readings* (pp. 253–273). Malden, MA: Blackwell.
- Biggar, H. (2005). NAEYC recommendations on screening and assessment of young English-language learners. *Young Children*, 60(6), 44–46.
- Bourassa, D. C., & Treiman, R. (2001). Spelling development and disability: The importance of linguistic

factors. *Language, Speech, and Hearing Services in Schools*, 32(3), 172–181.

CaMLA. (n.d.). YLTE. Retrieved from <http://www.cambridgemichigan.org/institutions/products-services/tests/proficiency-certification/ylte/>

CaMLA. (2014). YLTE 2014 Report. Ann Arbor, MI: CaMLA. Retrieved from <http://www.cambridgemichigan.org/wp-content/uploads/2015/02/YLTE-2014-Report.pdf>

Carless, D. (2012). *Young learners' perceptions of their assessment experiences: Achievement, anxiety and pressure*. Paper presented at the 34th Language Testing Research Colloquium, Princeton, New Jersey.

Carless, D., & Lam, R. (2014). The examined life: Perspectives of lower primary school students in Hong Kong. *Education 3–13: International Journal of Primary, Elementary, and Early Years Education*, 42(3), 313–329.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Carter, K. (1986). Test-wiseness for teachers and students. *Educational Measurement: Issues and Practice*, 5(4), 20–23.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, 19, 254–272.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307–331.

Common Core State Standards Initiative. (2010). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*. Retrieved from http://www.corestandards.org/wp-content/uploads/ELA_Standards1.pdf

Cook, J. L., & Cook, G. (2009). *Child development: Principles & perspectives* (2nd ed.). Boston, MA: Pearson.

Critten, S., Pine, K. J., & Messer, D. J. (2013). Revealing children's implicit spelling representations. *British Journal of Developmental Psychology*, 31(2), 198–211.

Dewey, D. P. (2004). A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts. *Studies in Second Language Acquisition*, 26(2), 303–327.

Field, J. (2009). The cognitive validity of the lecture-based question in the IELTS Listening paper. *IELTS Research Reports*, 9, 17–65. Retrieved from https://www.ielts.org/pdf/vol9_report1.pdf

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Erlbaum.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

Green, A. (2014). *Exploring language assessment and testing*. New York: Routledge.

Hall, K., J., Collins, S., Benjamin, S., Nind, M., & Sheehy, K. (2004). SATurated models of pupildom: Assessment and inclusion/exclusion. *British Educational Research Journal*, 30(6), 801–817.

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

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *The Canadian Modern Language Review*, 59(4), 567–587.

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

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Menken, K. (2008). *English language learners left behind: Standardized testing as language policy*. Clevedon: Multilingual Matters.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education & Macmillan.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational & Psychological Measurement*, 25(3), 707–726.

National Education Association. (n.d.). *Facts about Children's Literacy*. Retrieved from <https://www.nea.org/grants/facts-about-childrens-literacy.html>

Norris, J. M. (2008). *Validity evaluation in language assessment*. Frankfurt, Germany: Peter Lang.

Nunes, T., Bryant, P., & Bindman, M. (1997). Morphological spelling strategies: Developmental stages and processes. *Developmental Psychology*, 33(4), 637–649.

Paribakht, T. S., & Wesche, M. B. (1999). Reading and “incidental” L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language Acquisition*, 21(2), 195–224.

Pinker, S. (2011). *Words and rules: The ingredients of language*. New York: Harper Perennial.

Pukett, M. B., & Black, J. K. (2000). *Authentic assessment of the young child: Celebrating development and learning* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Pulido, D. (2004). The relationship between text comprehension and second language incidental vocabulary acquisition: A matter of topic familiarity? *Language Learning*, 54(3), 469–523.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three-and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, 59(2), 234–247.

Rogers, W. T., & Yang, P. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, 12(3), 247–259.

Sarnaki, R. E. (1979). An examination of test-wiseness in the cognitive domain. *Review of Educational Research*, 49(2), 252–279.

Schmidt, R. (2000). *Language policy and identity politics in the United States*. Philadelphia, PA: Temple University Press.

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2), 237–246. doi: 10.1177/1098214005283748

Treiman, R., & Cassar, M. (1997). Spelling acquisition in English. In C. A. Perfetti, L. Rieben, & M. Fayol (Eds.), *Learning to spell: Research, theory, and practice across languages* (pp. 61–80). Mahwah, NJ: Lawrence Erlbaum Associates.

Treiman, R., & Kessler, B. (2014). *How children learn to write words*. Oxford: Oxford University Press.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, 27(1), 33–52.

Weir, C. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave Macmillan.

Wheelock, A., Bebell, D. J., & Haney, W. (2000). What can student drawings tell us about high-stakes testing in Massachusetts? *Teachers College Record*, ID# 10634. Retrieved from <http://www.tcrecord.org/>

Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly*, 45(4), 628–660.

Winke, P., & Lim, H. (2014). The effects of testwiseness and test-taking anxiety on L2 listening test performance: A visual (eye-tracking) and attentional investigation. *IELTS Research Reports* (3), 1–30. Retrieved from <https://www.ielts.org/pdf/Winke%20and%20Lim.pdf>

Zdorenko, T., & Paradis, J. (2008). The acquisition of articles in child second language English: fluctuation, transfer or both? *Second Language Research*, 24(2), 227–250.

Zdorenko, T., & Paradis, J. (2012). Articles in child L2 English: When L1 and L2 acquisition meet at the interface. *First Language*, 32(1), 38–62.

Acknowledgements

We thank Changchang Yao and Michigan State University's Center for Language Teaching Advancement (CeLTA) for their help with this study. We also thank the child participants and their parents for working with us. We first presented results from this study at the Midwest Association of Language Testers (MwALT) annual conference in Ann Arbor, MI, in October 2014. This study was funded through the CaMLA (Cambridge Michigan Language Assessments) Spaan Research Grant Program 2014. We are grateful for their support. The views expressed in this paper are ours and not necessarily those of the funding institution. All mistakes are our own.

Appendix

Drawing

- Great! I like your picture! Pick out a sticker to put on your picture.
- Can you explain your drawing? What is it a drawing of? What is in this picture?
- Why did you want to draw this?
- What were you thinking while you were drawing?

General

- How was the test?
- How did the test make you feel?
- Which part was your favorite?
- Which part was your least favorite?

Difficulty

- Was the test difficult or easy?

Length

- Was it too long or too short? Was it okay?

Distraction

- Was it easy to concentrate on the test till the end?

Instructions

- Were the instructions clear? (showing the instructions)
- How were the pictures?

Concluding question

- Do you have any other comments about the test?




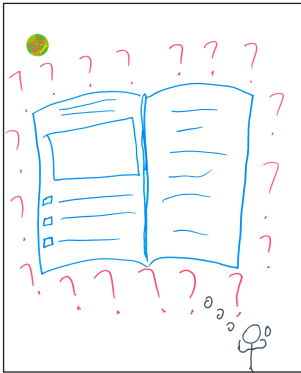


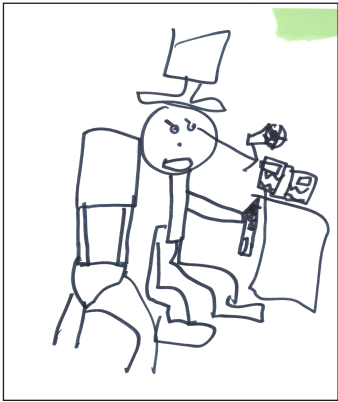
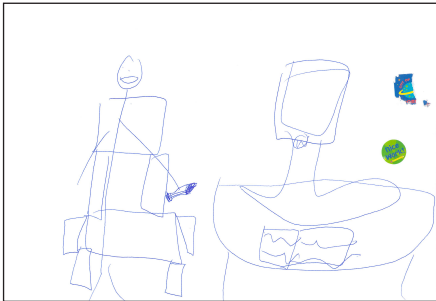
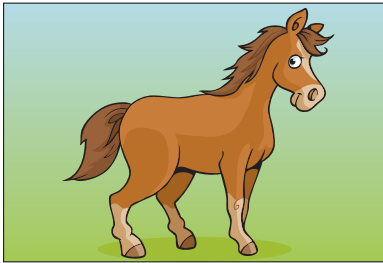
Participant	Bronze Drawing	Silver Drawing
#3 English speaker, age 8; coded as showing the Silver test more difficult.		
#11 Chinese speaker, age 9, coded as showing the Silver test more difficult.		
#13 Korean speaker, age 7; coded as showing no change in difficulty perceived.		
#14 Korean speaker, age 7, coded as showing the Bronze test more difficult (the only one coded in this direction).		

Figure 1. Examples of Bronze and Silver drawings. For all 37 drawings, email the authors.

A horse



I have four legs , two ears, two eyes, and long

(1) mane on my head. I'm a big animal. I don't live in

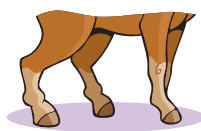
a (2) cage or a yard. I like eating

(3) hay and apples. I drink (4) water

A woman, a (5) Adult , or a child can ride me.

What am I? I am a horse.

example



legs



hippo



water



carrots



hair



man

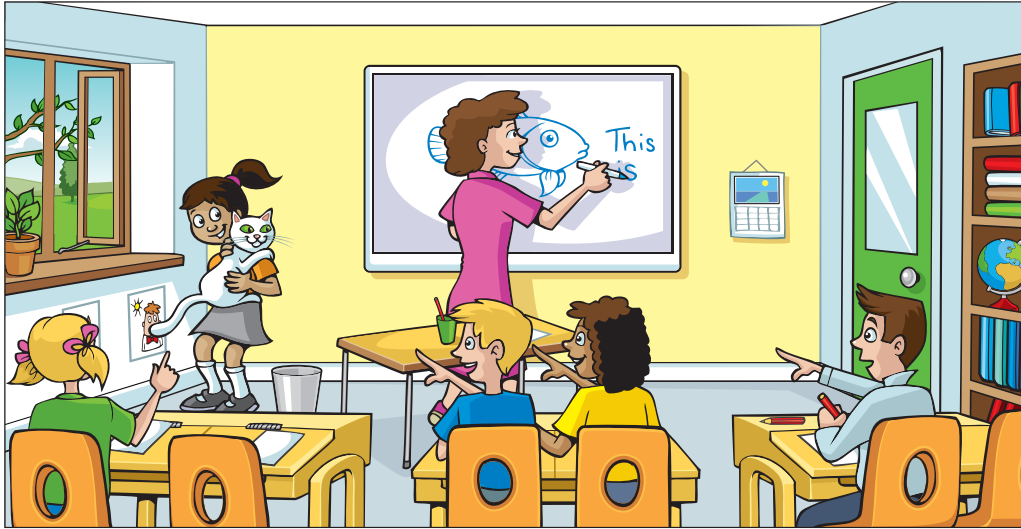


house

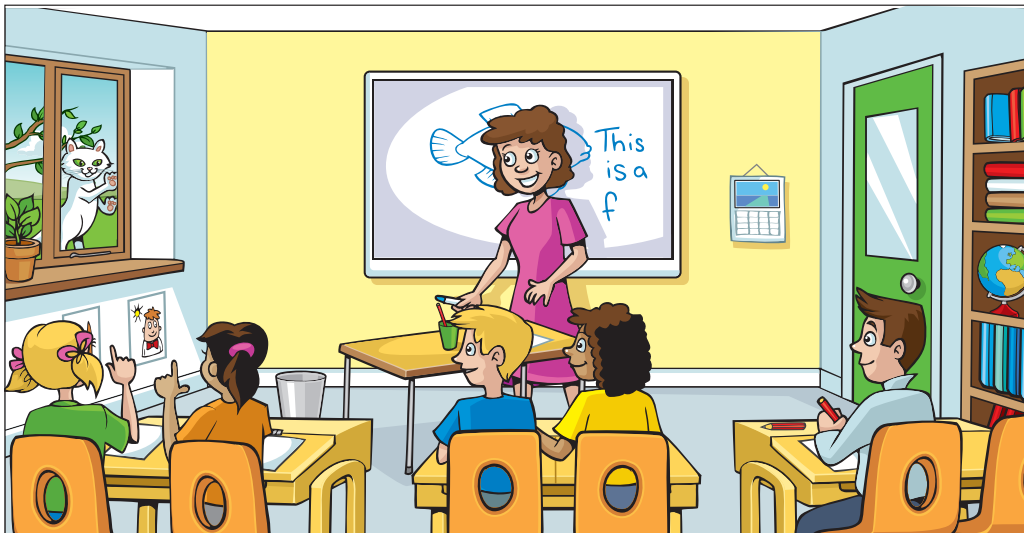


piano

Figure 2. An example of items 16 (number 1 in the test booklet) through 20 (number 5 in the test booklet) from the Bronze test (test taker #4). The directions, which the test takers read to themselves, stated, "Read this. Choose a word from the box. Write the correct word next to number 1–5. There is one example."



- 2 Who is holding the cat? a girl
- 3 What is the teacher doing now? looking



- 4 Where is the cat now? at the window
- 5 How many children are looking at the cat? five

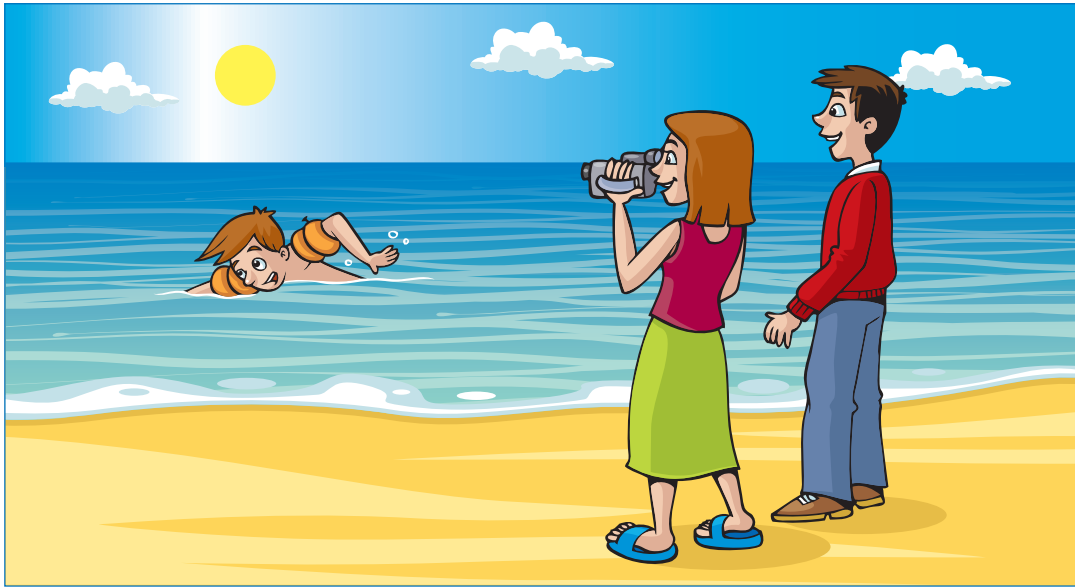
Figure 3. An example of items 22 (number 2 on the page) through 25 (number 5 on page) from the Bronze test (test taker #3). The directions, which the test takers read to themselves stated, "Look at the pictures and read the questions. Write one-word answers." No response box or word bank was provided; these were free-response items.



a bat

milk

Figure 4. An example of items 1 through 6 (numbers 1 to 6 on the page) from the Silver test (Participant 2). The directions, which the test takers read to themselves stated, "Look and read. Choose the correct words and write them on the lines. There is one example."



On Friday, the family ate breakfast in the garden because it was very sunny but Paul didn't want any. Then they all went to the beach again. The sea was very blue. Paul looked. There were three beautiful dolphins in the water! He ran to the sea and swam to them. Then Paul's dad threw a ball in the sea and the dolphins played with it. It was great and Paul stopped thinking about the sharks in the film. That evening, everyone in the family went to the movies again. This time the film was about a funny dolphin and they all enjoyed it.

7 The family had breakfast in garden on Friday.

8 Paul saw dolphin in the water.

Figure 5. An example of items 32 and 33 (numbers 7 and 8 on the page) from the Silver test (Participant 15).