*Feature Article*

# TRANSLANGUAGING IN READING COMPREHENSION ASSESSMENT: IMPLICATIONS ON ASSESSING LITERAL, INFERENTIAL, AND EVALUATIVE COMPREHENSION AMONG ESL ELEMENTARY STUDENTS IN TAIWAN

Curtis Shu-Sun Chu*
National Chung Cheng University

In an attempt to increase the accuracy of L2 reading assessments on the true knowledge of ESL learners, this study explores the three levels of reading comprehension (literal, inferential, and evaluative) and the notion of translanguaging for the possibility of an alternate testing method. Data were collected through a self-designed reading comprehension assessment among 123 sixth-graders in Taiwan. Two versions of the assessment were created: one provided reading text and questions in L2, and the other provided just the questions, translated into L1. Results indicate that student performance on all three comprehension levels were better with L1 questions and that all levels demonstrated high practical significance. Performance on evaluative open-ended questions was significantly better with L1 questions and L1-allowed responses. Correlations were found between students' L2 proficiency and comprehension, and on the inferential and evaluative levels, L2 proficiency had higher correlations with questions in L1 than in L2.

*Keywords: assessment, ESL, reading comprehension, TESOL, three levels, translanguaging*

**Contemporary** empirical studies have supported the recognition of reading comprehension as a multidimensional construct. Based on these studies, research has been initiated to compare and analyze reading comprehension tests on such aspects of the construct as the difference in processing demands (Kendeou, Papadopoulos, & Spanoudis, 2012), the kinds of skills being assessed (Keenan, Betjemann, & Olson, 2008), and the role of literal and higher order thinking in assessments (Basaraba, Yovanoff, Alonzo, & Tindal, 2013). The notion of the three levels of reading comprehension emerged—literal, inferential, and evaluative/critical comprehension (Herber, 1978). As we advance in understanding reading comprehension assessments, applying this knowledge growth to the context of second language (L2) acquisition could possibly expand our knowledge in the field.

While much research has been conducted to examine the connection between first-language (L1) and second- or foreign-language reading comprehension, testing practices in L2 remain problematic (García, 2009b; LaCelle-Peterson & Rivera, 1994; Short, 1993; Snow, 1998). In the modern era and its culturally diversified societies with their diverse population of English learners whose native language is not English, teachers have been challenged with the need to accurately and effectively assess what students really know. L2 learners were not assessed fairly for several reasons: (a) standardized assessments were normed for native speakers; (b) the items in assessments have lower reliability for limited English proficiency students; (c) students could have content knowledge but still be unable to comprehend test questions; and (d) students could have insufficient language proficiency to be assessed in English (The National Center for Fair and Open Testing, 2005). Assessing native speakers of Chinese or Spanish with an English reading comprehension assessment, for example, could fail to accurately capture what they really know.

In considering this problem, a seemingly simple question arose: What might happen if we employ L1 among L2 testing practices? This potential solution was inspired by the linguistic interdependence theory (Cummins, 1979) and the notion of translanguaging (García & Li, 2014). Unfortunately, there is only a handful of literature from the past four decades relating to alternative testing methods involving the use of L1 (Brantmeier, 2006; Gordon & Hanauer, 1995; Hock & Poh, 1979; Lee, 1986; Rahimi, 2007; Shohamy, 1984; Yu, 2008; Yun, Lee, & Park, 2012)—and among these studies, none examined the use of L1 and the three levels of reading comprehension assessment.

Thus, this study aims to be one of the first to explore the effects of integrating L1 in L2 reading comprehension assessment, with particular focus on the three levels of comprehension. Using the self-designed Elementary Level English Reading Comprehension Assessment (ELERCA), it compared student performance among sixth-grade ESL learners in Taiwan. Two versions of the same assessment were produced. Both versions provided the same reading texts in L2 (English), and employed multiple-choice and open-ended questions that were designed according to the three levels of reading comprehension. The only difference between the two versions was that one version provided translated questions and multiple-choice options in L1 (Chinese).

## What Do Reading Comprehension Tests Assess?

Reading tests existed long before the development of our modern concepts on reading. Without the benefit of contemporary scientific evidence, constructs of reading tests were developed based on intuition (Alderson, 2000). As studies on reading comprehension emerged, however, researchers had the necessary theories and frameworks to refer to and explore what reading tests really assessed. The reading tests cited in this section were mostly concerned with children learning to read.

In a recent study, three reading comprehension assessments—the Woodcock-Johnson Passage Comprehension (WJPC), a curriculum-based measure test, and a recall test—were compared and analyzed for their processing demands (Kendeou et al., 2012). The study discovered that one test required significant orthographic process and working memory skills; another required fluency and vocabulary; and the third required phonological process, orthographic process, and working memory skills. Another, similar study compared four reading comprehension assessments—the WJPC, the Gray Oral Reading Test (GORT), retellings and comprehension questions from the Qualitative Reading Inventory (QRI), and the Peabody Individual Achievement Test (PIAT) (Keenan et al., 2008). The study found that the PIAT and WJPC were highly dependent on decoding instead of listening comprehension, whereas the GORT and QRI were highly dependent on listening comprehension instead of decoding. This implies that students with adequate decoding skills but poor listening comprehension skills were likely to perform poorly on the GORT or QRI, and might be labeled as slow readers.

A subsequent study comprising approximately 2,400 students filled the gap on the lack of research on reading comprehension by examining the Rasch (RIT) scale difficulties on the three levels of reading comprehension (Basaraba et al., 2013). Even though the study discovered that literal items were significantly less challenging compared to inferential and evaluative items, the relationship among the three levels was nonlinear; this means that high performance on any single level of comprehension does not positively relate to performance on the other two levels of comprehension. Thus, it seems naïve to label students as struggling or advanced readers based simply on results from reading comprehension assessments. Evidence has proven that assessments do not necessarily measure the same skills, and could also measure different aspects, such as the three levels of comprehension.

## The Three Levels of Reading Comprehension

As noted, the notion of the three levels of reading comprehension, also known as the three-level guide, was initially proposed by Herber (1970) and further developed by Vacca and Vacca (1999). The three levels consist of literal, interpretative/inferential, and applied/evaluative comprehension. Literal comprehension refers to the ability to comprehend the text at word level, reading the lines, and searching for literal information. Interpretive/Inferential comprehension refers to the ability to read between the lines, identify relationships among information, and comprehend the author's intended meaning. Finally, applied/evaluative comprehension refers to the ability to read beyond the lines, engage extensive background knowledge, and apply/evaluate information. The three levels were initially thought to be hierarchical, but in 1978, Herber recast his idea and deemed that the three levels could be operated both top-down or bottom-up.

The three levels could be better understood as a reference to a similar notion of Herber's (Herber, 1970), which is the taxonomy proposed by Pearson and Johnson (1978). The taxonomy placed readers' responses into three categories: textually explicit, textually implicit, and scripturally implicit. Textually explicit responses are found directly from the text (literal); textually implicit responses require logical inference for the justification of the responses in addition to the criteria of textually explicit responses (inferential); and scripturally implicit responses require readers to activate their prior knowledge because only the question could be found in the text (evaluative). It is crucial to recognize that reading comprehension, particularly comprehension requiring higher order thinking skills, is influenced by prior knowledge (Cain, Oakhill, Barnes, & Bryant, 2001; Kintsch, 1988).

The roots of literal, inferential, and evaluative comprehension skills might have stemmed from *Bloom's Taxonomy* (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). Bloom classified thinking behavior into three domains: affective, psychomotor, and cognitive. Among these, a continuum from lower to higher order thinking skills was identified in the cognitive domain. The continuum ranges from knowledge, comprehension, application, and analysis to synthesis and evaluation. *Bloom's Taxonomy* was later revised (Anderson et al., 2001), and the revised continuum in the cognitive domain encompasses lower to higher order thinking skills, including remembering, understanding, applying, analyzing, evaluating, and creating. While textbooks and studies related to the three levels of comprehension among classrooms exist (Cooter & Flynt, 1996; Vacca & Vacca, 1999), no literature was found on the application of the three levels in the field of second language acquisition. Thus, in this study I sought to justify the feasibility of exploring the three levels in the field of ESL with Cummins' (1979) linguistic interdependence theory and García's translanguaging (2009b) theory.

## Translanguaging and the Language Interdependence Theory

The linguistic interdependence theory proposed by Cummins (1979) posited that bilinguals do not separately store two languages. Instead, there is a cognitive interdependence known as the common underlying proficiency—translanguaging. The assumption is that regardless of the difference in languages, proficiencies that require higher and cognitively demanding skills should be common across languages. Recent studies have extended Cummins' theory and discovered that while one language is being used, the other language would not remain dormant but instead becomes activated (Hoshino & Thierry, 2011, 2012). In other words, because inferential and evaluative comprehension involves higher order thinking skills, such skills should be common across languages.

Cummins' theory played a crucial part in García's (2009b) concept of translanguaging, but before defining this concept, understanding García's argument for the term "emergent bilinguals" would allow us to visit the perspective of translanguaging. Instead of labeling speakers of a language other than English as English language learners or limited English proficient students, García (2009a) argued for the term as more appropriate to indicate how education for bilingual students was generally misunderstood across the globe. With particular reference to the United States, students who are not native English speakers and are English learners were often neglected because of their apparent bilingual ability. The possibility to further develop their bilingualism during schooling was also neglected. García argued that the term "emergent bilinguals" would enable viewing bilingualism as a potential resource, rather than as a limitation or problem when compared with the abilities of speakers of English.

Given this environment, to maximize communicative potential, bilinguals access different linguistic features or autonomous languages through the act of translanguaging (García & Li, 2014). Translanguaging allows speakers to access their full linguistic repertoire without having to be constantly aware of socially and politically defined boundaries of named, national, and state languages (Otheguy, García, & Reid, 2015). This flexible use of linguistic resources liberates the voices of linguistic minority students, promotes deep and reflexive thinking, and enables rigorous cognitive engagement with texts. In addition, students could better comprehend difficult texts when they can access their background knowledge in other languages (García, 2014). If we were to assess how well they comprehend a L2 text on the three levels of comprehension, students should be allowed to access their entire linguistic repertoire in order for them to fully demonstrate their comprehension skills, particularly on the inferential and evaluative higher order thinking levels. Based on the linguistic interdependence theory and the notion of translanguaging, it would seem logical to assess reading comprehension by asking questions and allowing responses in students' L1. Yet, only a handful of studies examining the use of L1 in L2 testing could be found.

## Previous Studies with the Use of L1 in L2 Testing

As early as 1979, a study on language assessment among ESL students in Malaysia discovered that questions written in L2 might have a negative effect on learners at their early stages of English acquisition (Hock & Poh, 1979). Subsequently, Shohamy (1984) conducted a pioneer study on the use of questions written in L1. This study looked into the effects of providing multiple-choice and open-ended questions in both L1 and L2 upon reading the same text in L2. The results showed that both kinds of questions provided in L1 were easier than in L2 and that the effect was strongest on low-level students. Shohamy speculated on several possible reasons: (a) L1 questions might reduce anxiety among low-level learners; (b) L1 wordings might have provided hints; (c) questions in L2 might increase exposure to unfamiliar vocabulary; and (d) L1 questions might be thought of as more authentic.

Gordon and Hanauer (1995) conducted similar research by testing 28 EFL learners in two Israeli public high schools. Adopting a reading comprehension test with an expository text and comprehension questions, the study provided multiple-choice and open-ended questions in both Hebrew and English. With similar findings as Shohamy's, multiple-choice questions provided in L1 were found to be the easiest; open-ended questions provided in L2 were the most difficult. Gordon and Hanauer concluded that when testing tasks are provided and written in L1, there would be a greater potential for open-ended questions to become an information source for test takers.

Aside from multiple-choice or open-ended testing methods, a study was conducted on the comparison of a recall task written in L1 and L2 shortly after Shohamy's pioneer study (Lee, 1986). In the study, Lee discovered that learners from beginning to intermediate levels of language instruction achieved a significantly higher written score when they were allowed to write in their native language, compared to those who wrote in L2. Two decades later, the findings of both Brantmeier (2006) and Rahimi (2007) further conform with Lee's, reinforcing that readers with low L2 proficiency perform better when the test method provides questions in their L1.

In the only study involving the Chinese language, 157 Chinese undergraduates wrote summaries of English texts in both Chinese and English (Yu, 2008). The study discovered that although students wrote significantly longer summaries in Chinese, the quality of their summaries was consistently poorer when compared to English summaries; in spite of these results, Yu concluded that Chinese summaries were better for measuring English reading comprehension. Finally, a more recent study conducted in Korea confirmed the influence of L1 and L2 in terms of instructions, questions, and choices (Yun et al., 2012). The study used the College Scholastic Ability Test to collect data from 641 second-grade high school students. While student performance on listening comprehension was not affected regardless of providing L1 or L2 instructions, students who took the English test with L1 instructions, questions, and choices outperformed those who took the test entirely in English.

Unfortunately, none of the studies explored students in the age range of elementary to junior high school, nor did any of the studies examine the relationship between the use of L1 in L2 testing and reading comprehension skills. Therefore, this study aims to explore the use of L1 in L2 testing and its influence on the three levels of comprehension through the following questions:

1. In an English reading comprehension assessment, could multiple-choice questions and options provided in L1 (Chinese) more accurately assess students' knowledge compared to assessment entirely in L2 (English)?
2. Are students' abilities to make inferences and perform evaluative/critical thinking underestimated in traditional assessment done entirely in L2?
3. Are open-ended evaluative questions provided in L1 and allowing L1 responses more effective in terms of assessing students' evaluative/critical thinking ability?
4. How does the students' English proficiency level affect their performance on questions provided in L1 and L2?

## Method

This study employed the self-designed Elementary Level English Reading Comprehension Assessment (ELERCA) as the tool for data collection. There are four reading passages in ELERCA: a sample student worksheet, a passage of three self-introductions, a letter to Santa, and a party flyer. The passages range from 92 to 190 words, and there are seven multiple-choice questions (providing three options) following each passage. Among the seven questions, there were two literal, two inferential, and three evaluative questions (including one open-ended question). ELERCA has a total of 24 multiple-choice questions and four open-ended questions. The L2 (English) version of ELERCA was first designed, and then the questions and options were translated for the L1 (Chinese) version. Both versions provided the same reading text in L2, but one had the questions and options in L1 while the other remained entirely in L2. A frequency check of Chinese words using the Academia Sinica Institute of Linguistics online database was performed to avoid adapting low-frequency words upon translation.

Literal, inferential, and evaluative/critical comprehension questions were designed in reference to the literature and consideration of engaging students' prior knowledge at the evaluative level. Integrating the format and criteria above, a definition for the three levels of comprehension was summarized and referred to throughout the design of ELERCA.

After filtering for ineffective samples, 123 sixth-grade elementary students in five classes of a public school in Taiwan participated in this study. The reading comprehension assessment was administered during their compulsory English classes. Through random sampling, data were obtained for three classes (65 students) writing the Chinese (L1) version of the test, and another three classes (58 students) writing the English (L2) version. Results are shown in Table 1.

**Table 1** Summarized Definition and Sample Questions of the Three Levels of Comprehension

| Three Levels | Summarized Definition | Sample Questions, Options, and Correct Answers |
|---|---|---|
| Literal | Questions should assess text-based comprehension. Answers to the questions could be found directly in the text of their corresponding passages. | What does Andy want for Christmas?<br>A. He wants a computer.<br>B. He wants a bicycle.<br>C. He wants a dog (as stated in the text). |
| Inferential | Questions should assess students' ability to make inferences, thus requiring students to go beyond the facts from passages and make inferences about meanings that are not explicitly stated in the text. | Who is the oldest?<br>A. Andy is the oldest.<br>B. Bill is the oldest.<br>C. Lisa is the oldest (infer from explicit clues in text). |
| Evaluative | The most complex level of the three types of reading comprehension, these questions should assess students' ability to evaluate and analyze information obtained from the text. Evaluative comprehension requires more understanding of the text (literal), making interpretations on the meaning of the text (inferential), and evaluating the information of the text with greater integration of prior knowledge or experiences. | Which is the best reason for Bill to want a new ball?<br>A. His friends asked him to get a new ball.<br>B. He and his friends have no ball to play with.<br>C. He can practice on weekdays (text-explicit: Nine-year-old Bill will play ball with his friends on weekends. Thus, evaluating this information, Bill and his friends would already have a ball to play with, and on weekdays he could practice to prepare for weekends, thus out ruling options A and B). |

Each question had only one correct answer. As for the four open-ended evaluative questions, they were graded in reference to Table 1. If the participant's response demonstrated a high level of evaluative thinking, it will result in a score of 1; if only partial evaluative thinking was demonstrated, it will result in a score of 0.5. The four evaluative open-ended questions were graded among three teachers, and we reached consensus for every response.

The reading passages in ELERCA were written by referring to the simulation tests of the General English Proficiency Test for Kids (GEPT Kids). GEPT Kids was developed by The Language Training & Testing Center (Lee, 2015a, 2015b) specifically for elementary students in Taiwan and was approved by the Taiwan Ministry of Education (MOE). Highly valued in Taiwanese society, it was administered regularly throughout the year. In addition, to assure that the difficulty of the assessment suits the level of elementary students, vocabularies in ELERCA were chosen from the MOE Elementary School 300 Fundamental Vocabularies (MOE, 2001) and the MOE Junior High School and Elementary School 1200 Most Fundamental Vocabularies (MOE, 2008).

In the study, the reliability of ELERCA is determined by the Cronbach $\alpha$, which measures the internal consistency of the assessment. The Cronbach $\alpha$ of the English version was .860 and the Chinese version was .80, therefore indicating a high level of internal consistency for both versions. In terms of validity, as suggested by Lynn (1986), nine experts were consulted, including schoolteachers in an education doctoral program and professors with expertise in reading comprehension. They were asked to rate the appropriateness of the initially designed 36 questions on a scale from 1 to 10 and provide written suggestions before finalizing the testing version. The content validity index (CVI) for each of the questions was calculated to maintain a high level of relevance, and questions with a CVI of .80 or higher were adopted (Lynn, 1986). As for questions with a CVI below 0.80, they were either removed or modified according to the suggestions of the experts. The overall CVI of the testing version was 0.91, indicating a very high level of content validity.

### Data Analysis

Statistical analyses were performed using SPSS 17.0 software to help answer the assumptions of this study. First, t-tests were conducted to compare students' English and Chinese proficiency level. This was done to ensure that the two groups who took different versions of ELERCA were comparable and without significant differences in language proficiency levels. To compare the performance of the two groups in detail, t-tests were again conducted, including their performance on each of the three levels of comprehension; their performances then were compared to reveal differences for taking the L1 or L2 version. Finally, a correlation analysis was conducted to determine the relationships between students' English proficiency level and their performance on the three levels.

## Results

Descriptive statistics of student performance on both L1 and L2 versions of ELERCA is shown in Table 2. Literal, inferential, evaluative (with open-ended responses), and the open-ended responses were separately analyzed. Students achieved a higher total score on all levels of comprehension in the L1 version as well.

**Table 2** Descriptive Analysis of Student Performance on Both Versions (L1 and L2) of ELERCA

| | *n* | | *M* | | *SD* | |
|---|---|---|---|---|---|---|
| Student performance | L1 | L2 | L1 | L2 | L1 | L2 |
| Total score | 65 | 58 | 19.34 | 12.87 | 5.07 | 6.06 |
| Literal questions (8) | 65 | 58 | 5.85 | 4.53 | 2.12 | 2.33 |
| Inferential questions (8) | 65 | 58 | 5.98 | 4.22 | 1.53 | 2.09 |
| Evaluation questions (12) | 65 | 58 | 7.51 | 4.11 | 2.44 | 2.50 |
| Open-ended responses (4) | 65 | 58 | 1.85 | .595 | 1.33 | .91 |

### Language Proficiency among Participants

First, to ensure that students who took the L1 version and those who took the L2 version were comparable, t-tests were conducted to compare their language proficiency, as shown in Table 3. Their language proficiency was calculated from their performance on their Chinese and English midterm and final exams when they were fifth-graders. No significant differences were found in terms of which version of the test they took or their English/Chinese proficiency. Therefore, the two groups who took the L1 or L2 test had similar language proficiency, thus demonstrating that the two groups were comparable.

**Table 3** Participants' Language Proficiency

| | English Proficiency Score | | | | |
|---|---|---|---|---|---|
| | *n* | *M* | *SD* | *t* | *p* |
| Took L1 test | 65 | 86.277 | 20.0511 | –.898 | .371 |
| Took L2 test | 58 | 89.224 | 15.7937 | | |
| | **Chinese Proficiency Score** | | | | |
| | *n* | *M* | *SD* | *t* | *p* |
| Took L1 test | 65 | 84.415 | 9.836 | –.630 | .530 |
| Took L2 test | 58 | 85.448 | 8.1417 | | |

*p*<.01

## Reading Comprehension Performance of L1 and L2 Test Takers for the Three Levels of Comprehension

Referring to the first research question, whether multiple-choice questions and options provided in L1 (Chinese) could more accurately assess students' knowledge compared to assessment entirely in L2 (English), t-tests were conducted to compare students' performance to determine whether questions provided in L1 could more accurately assess students' knowledge. A significant difference was found when comparing the means of participants' total score among the two groups who took the L1 and L2 versions of ELERCA, as shown in Table 4. Students who took the L1 test (19.3385) significantly outperformed those who took the L2 test (12.8707). In addition, the Cohen's effect size value (*d* = 1.16) suggested a high practical significance (Cohen, 1988). In other words, the average score of students taking the L1 version of ELERCA was higher than those taking the L2 version by a 1.16 standard deviation. When the Cohen's effect size value is converted into Cohen's U3, it could be interpreted that 87.7% of the L1 test takers will achieve a score above the average of the L2 test takers. When the Cohen's effect size is converted into common-language effect size (Ruscio, 2008), it could be interpreted as a 79.4% chance for L1 test takers to perform better than L2 test takers.

**Table 4** Results of T-Tests for Reading Comprehension Performance

| | *n* | *M* | *SD* | *t* | *d* | *p* |
|---|---|---|---|---|---|---|
| Took L1 test | 65 | 19.3385 | 6.05788 | 6.444 | 1.16 | .000 |
| Took L2 test | 58 | 12.8707 | 5.06837 | | | |

*p*<.01

As for the second research question, whether students' abilities to make inferences and perform evaluative/critical thinking were underestimated, results from t-test analysis found significant differences on all three levels when comparing the mean scores of the two groups. On the literal level, the mean score between the two groups differs by only 1.31 points, as shown in Table 5. Students who took the L1 test (5.8462) significantly outperformed those who took the L2 test (4.5345), and the Cohen's effect size value (*d* = 0.58) suggests a moderate practical significance. This could be interpreted as students who took the L1 version had a higher score than those taking the L2 version by a 0.58 standard deviation. Of the test takers, 71.9% will achieve a score above the average of the L2 test takers, and there is a 65.9% chance for L1 test takers to perform better than L2 test takers on the literal level.

**Table 5** Results of T-Tests for Performance on the Three Levels of Comprehension

|  | Literal Comprehension |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  | *n* | *M* | *SD* | *t* | *d* | *p* |
| Took L1 test | 65 | 5.8462 | 2.12302 | 3.264 | 0.58 | .001 |
| Took L2 test | 58 | 4.5345 | 2.33370 |  |  |  |
|  | Inferential Comprehension |  |  |  |  |  |
|  | *n* | *M* | *SD* | *t* | *d* | *p* |
| Took L1 test | 65 | 5.9846 | 1.52574 | 5.273 | 0.96 | .000 |
| Took L2 test | 58 | 4.2241 | 2.09463 |  |  |  |
|  | Evaluative Comprehension<br>(with open-ended questions) |  |  |  |  |  |
|  | *n* | *M* | *SD* | *t* | *d* | *p* |
| Took L1 test | 65 | 7.5077 | 2.43588 | 7.613 | 1.37 | .000 |
| Took L2 test | 58 | 4.1121 | 2.50621 |  |  |  |
|  | Evaluative Comprehension<br>(open-ended questions only) |  |  |  |  |  |
|  | *n* | *M* | *SD* | *t* | *d* | *p* |
| Took L1 test | 65 | 1.8462 | 1.32854 | 6.147 | 1.10 | .000 |
| Took L2 test | 58 | .5948 | .91026 |  |  |  |

*p*<.01

On the second level, then, the students who took the L1 test (5.9846) performed significantly better than those who took the L2 test (4.2241), and Cohen's effect size value (*d* = 0.96) suggests a high practical significance. This means that students taking the L1 version had a higher score than those taking the L2 version by a 0.96 standard deviation. Of the test takers, 83.1% will achieve a score above the average of the L2 test takers, and there is a 75.1% chance for L1 test takers to perform better than L2 test takers on the inferential level.

Among the three levels, students' performance was the most significant on the third level, the evaluative—whether open-ended evaluative questions provided in L1 and allowing L1 responses could better capture students' evaluative/critical thinking ability. When open-ended evaluative questions were included, students who took the L1 test (7.5077) significantly outperformed those who took the L2 test (4.1121), and the Cohen's effect size value (*d* = 1.37) suggests a high practical significance. This result implies that students taking the L1 version had a higher score than those taking the L2 version by 1.37 standard deviation. Of the test takers, 91.5% will achieve a score above the average of the L2 test takers, and there is an 83.4% chance for L1 test takers to perform better than L2 test takers on the evaluative level.

On the third level, students who took the L1 test (1.8426) significantly outperformed those who took the L2 test (.5948). The Cohen's effect size value (*d* = 1.10) suggests a high practical significance. As an explanation, the average score of students taking the L1 version is higher than those taking the L2 version by 1.10 standard deviation. Of the L1 test takers, 86.4% will achieve a score above the average of the L2 test takers, and there is a 78.2% chance for L1 test takers to perform better than L2 test takers on open-ended evaluative questions. In addition, students were nearly four times as likely to attempt to write their responses in L1. Only nine students scored zero on a total of four open-ended questions in the L1 test, whereas 35 students scored zero all four open-ended questions in the L2 test. Thus, students were more likely to respond and perform significantly better when open-ended questions were provided in L1 and allowing L1 responses.

### Correlations Between Language Proficiency and Reading Comprehension Performance

In response to the final research on how might students' language proficiency affect their performance on questions provided in L1 or L2, correlation analyses were performed among students' English proficiency and their performance. Significance was found between students' language proficiency and reading comprehension total score, as shown in Table 6. Students' English proficiency was moderately correlated with total reading comprehension score on both L1 and L2 questions. Their Chinese proficiency is moderately correlated with total reading comprehension score in L2 and highly correlated in L1; in other words, students' English (.561**>.490**) and Chinese proficiency (.647**>.489**) were higher correlated to their total performance on L2 compared to their performance on L1 questions. It was expected that students' Chinese proficiency should have a higher correlation when responding to questions in their native language, but it was unexpected that their English proficiency had a higher correlation as well.

**Table 6** Correlation of Participants' Language Proficiency and Total Score

| | | | | *r* | |
|---|---|---|---|---|---|
| | *n* | *M* | *SD* | English Proficiency | Chinese Proficiency |
| L1 testing | 65 | 19.3385 | 5.06837 | .561** | .647** |
| L2 testing | 58 | 12.8707 | 6.05788 | .490** | .489** |

*p*<.01

### English Proficiency and the Three Levels of Comprehension

First, significance was found between students' English proficiency and their performance on all three levels, as shown in Table 7. It was expected that students' English proficiency was moderately correlated with literal comprehension on both L1 and L2 versions (.545**<.548**). This finding matches the logical assumption that the more proficient a student is in English, the better that student would perform on reading assessments.

**Table 7** Correlation of Participants' English Proficiency and Three Levels of Comprehension

| | Literal | | | |
|---|---|---|---|---|
| | *n* | *M* | *SD* | *r* |
| L1 testing | 65 | 5.8462 | 2.12302 | .545** |
| L2 testing | 58 | 4.5345 | 2.33370 | .548** |
| | Inferential | | | |
| | *n* | *M* | *SD* | *r* |
| L1 testing | 65 | 5.9846 | 1.52574 | .494** |
| L2 testing | 58 | 4.2241 | 2.09463 | .365** |
| | Evaluative (with open-ended questions) | | | |
| | *n* | *M* | *SD* | *r* |
| L1 testing | 65 | 7.5077 | 2.43588 | .382** |
| L2 testing | 58 | 4.1121 | 2.50621 | .367** |
| | Evaluative (open-ended questions only) | | | |
| | *n* | *M* | *SD* | *r* |
| L1 testing | 65 | .5948 | .91026 | .507** |
| L2 testing | 58 | 1.8462 | 1.32854 | .369** |

*p*<.01

Second, students' English proficiency was moderately correlated with inferential comprehension on the L1 version, but only modestly correlated in L2. The correlation between students' English proficiency and their performance on L1 inferential questions not only was higher than the correlation with questions in L2 (.494**>.365**), but also was the highest among the three levels of comprehension. In particular, while the positive relationship between English proficiency and performance on inferential comprehension exists, it was unexpected that this relationship was even more significant when students were being tested in L1 rather than the L2 version.

Finally, students' English proficiency was only modestly correlated with evaluative comprehension on both the L1 and L2 versions. The correlation between students' English proficiency and their evaluative comprehension on the L1 version was slightly higher than the correlation with their performance on the L2 version (.382**>.367**). If considering only the evaluative open-ended questions, however, it would be a very different situation. Students' English proficiency was moderately correlated with open-ended evaluative comprehension on the L1 version, and only modestly correlated with the performance on the L2 version. When open-ended evaluative questions were provided in L1 and responses in L1 were allowed, students' English proficiency and their performance had a much higher correlation than providing questions in L2 and allowing L2 only responses (.507**>.369**).

## Discussion

Findings in this study support its first assumption: When compared to traditional assessment entirely in L2, comprehension assessments with L2 text and questions provided in L1 could more accurately assess students' knowledge. Because the aim of more accurately assessing students' knowledge can be rather abstract and complicated, however, it is difficult to prove simply based on the finding that students taking the L1 test significantly outperformed those who took the L2 version. This finding confirmed previous studies but with slightly different situations—i.e., questions in L1 were easier than in L2, L2 open-ended questions were the most difficult, and students outperforming those who were tested entirely in L2 (Gordon & Hanauer, 1995; Shohamy, 1984; Yun et al., 2012). Therefore, with no previous literature present, this study aimed to discover further evidence to support the first assumption by examining assumptions two, three, and four.

Assumption two of this study aimed to examine whether students' abilities to make inferences and perform evaluative/critical thinking were underestimated in traditional assessments made entirely in L2. When the L1 version was compared to the L2 version, significant differences were found on all three levels. There was also a 0.96 standard-deviation increase on their inferential scores and a 1.37 standard-deviation increase on their evaluative scores. In addition, students' English proficiency has a stronger correlation with responding to questions in L1 instead of in L2. Considering Cummins' (1979) theory of linguistic interdependence, it might be possible that questions in L1 enable students to better demonstrate their common underlying proficiency; in other words, questions in L1 might better capture students' inferential and evaluative comprehension, which involves higher order thinking. Thus, such findings might indicate the possibility of undermining students' inferential and evaluative comprehension skills in reading assessments that are entirely in L2.

Assumption three anticipated that open-ended evaluative questions in L2 and requiring responses in L2—compared to questions in L1 and allowing responses in L1—to be less effective on assessing students' evaluative/critical thinking ability. Students' performance on open-ended evaluative questions significantly increased by a 1.10 standard deviation when the questions were provided in L1. In addition, students' response rates were nearly four times higher when they were allowed to respond in L1. This phenomenon could be explained by the notion of translanguaging. Allowing students to respond in their native language enables them to access their full linguistic repertoire without the constraint of having to be constantly aware of their L2 (Otheguy et al., 2015). Allowing responses in L1 could also promote reflexive thinking and enable access to background knowledge in other languages (García, 2014), which are elements of higher order thinking. Thus, through the act of translanguaging, students might become more capable or comfortable to demonstrate their higher order thinking in their native language, which might be why their performance significantly increased.

The fourth, and final, assumption presumed that students' English proficiency would not affect their performance regardless of providing questions in L1 or L2. As no literature examined students' language proficiency and their performance on the three levels of comprehension, this study acts as one of the first in this regard. Students' English proficiency was discovered to have higher correlations on both inferential and evaluative levels when responding to questions in L1 rather than responding in L2. The correlation between English proficiency and literal comprehension was the highest among the three levels regardless of providing questions in L1 or L2. This finding was expected, because the literal level demands comprehension at the word level, which seemed to match the general English education environment at the elementary level, with its emphasis on vocabulary and sentence structure. On the inferential level, it was surprising to discover that the correlation with English proficiency and the responses to L1 questions was much higher than the correlation with the responses to L2 questions. In other words, when questions involving inferential comprehension in a reading assessment were provided in L1, it is more likely for students with high English proficiency to perform well compared to when questions are provided in L2. This statement applies to the evaluative level as well, particularly on evaluative open-ended questions. Therefore, the implication is that students' English proficiency is better reflected on the inferential and evaluative level in reading comprehension assessments when questions, options, and responses for open-ended questions are in L1.

## Conclusion

Testing practices are problematic in L2, as they could not accurately capture what students really know and perhaps undermine students' inferential and evaluative comprehension skills. The translanguaging reading comprehension assessment method proposed in this study implies the possibility of capturing a more accurate understanding of students' true knowledge. Integrating the notion of translanguaging into reading comprehension assessment on the three levels of comprehension, two versions of the self-designed ELERCA were employed in this study. ESL learners performed significantly better when they were provided with L2 reading texts having multiple-choice questions, options, and allowed responses in L1 on open-ended questions. They performed significantly better on all three levels of comprehension, and their English proficiency was surprisingly higher correlated with responding to inferential and evaluative questions in Chinese than in English. The translanguaging reading comprehension assessment method not only enables students to access their full linguistic repertoire, but also promotes reflexive thinking and facilitates access to background knowledge in other languages, crucial elements of higher order thinking. The method also enables students to demonstrate their common underlying proficiency. Thus, students might become more capable and comfortable in demonstrating their higher order thinking when their native language is involved in reading comprehension assessments.

The assessment method described in this study could be particularly useful in a time seeing an increasingly diverse population of English learners whose native language is not English. In the classroom, it could be applied when teachers believe that students know something about a certain topic but one they do not know how to express in English. Specifically, when the purpose of a reading comprehension assessment is to determine how well L2 (English) learners comprehend the text, this assessment method would be suitable. Limiting responses in L2 and requiring decoding of both L2 questions and multiple-choice options could interfere with leaners' expression of what they really know. In addition, compared to traditional reading comprehension assessment entirely in L2, this assessment method could better capture L2 learners' higher order thinking skills—namely, the ability to make inferences and critically evaluate presented texts.

Regarding the limitations of this study, more samples could be collected from different parts of Taiwan in the future. As one of the first studies to explore the elementary level and the three levels of comprehension among ESL students, more variables should be designed for analysis. At the same time, variables could be introduced and controlled to produce a research design even more rigorous from the perspective of experimental study. In addition, cross-cultural comparison with second language learners in languages other than Chinese or English would provide further evidence on the effects of integrating the notion of translanguaging in reading comprehension assessments.

## References

Alderson, J. C. (2000). *Assessing reading.* Cambridge, UK: Cambridge University Press.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., . . . Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of* Bloom's taxonomy of educational objectives. New York, NY: Pearson, Allyn & Bacon.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing, 26*(3), 349–379.

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, Handbook I: The cognitive domain.* New York, NY: David McKay.

Brantmeier, C. (2006). The effects of language of assessment and L2 reading performance on advanced readers' recall. *The Reading Matrix, 6*(1).

Cain, K., Oakhill, J., Barnes, M. A., & Bryant, P. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition, 29,* 850–859.

Cooter, R. B., & Flynt, E. S. (1996). *Teaching reading in the content areas: Developing content literacy for all students*. Englewood Cliffs, NJ: Prentice Hall.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children.

*Review of Educational Research, 49*(2), 222–251.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

García, O. (2009a). Emergent bilinguals and TESOL: What's in a name? *TESOL Quarterly, 43*(2), 322–326.

García, O. (2009b). Education, multilingualism and translanguaging in the 21st century. In A. Mohanty, M. Panda, R. Phillipson, & T. Skutnabb-Kangas (Eds.), *Multilingual education for social Justice: Globalising the local* (pp. 128–145). New Delhi, India: Orient Blackswan.

García, O. (2014). TESOL Translanguaged in NYS: Alternative perspectives. *NYS TESOL Journal, 1*(1), 2–10.

García, O., & Li, W. (2014). *Translanguaging: Language, bilingualism and education.* Basingstoke, UK: Palgrave Macmillan.

Gordon, C. M., & Hanauer, D. (1995). The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly, 29*(2), 299–324.

Herber, H. L. (1970). *Teaching reading in content areas*. Englewood Cliffs, NJ: Prentice Hall.

Herber, H. L. (1978). *Teaching reading in content areas* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Hock, T. S., & Poh, L. C. (1979). The performance of a group of Malay-medium students in an English reading comprehension test. *RELC Journal, 10,* 81–89.

Hoshino, N., & Thierry, G. (2011). Language selection in bilingual word production: Electrophysiological evidence for cross-language competition. *Brain Research, 1371,* 100–109.

Hoshino, N., & Thierry, G. (2012). Do Spanish-English bilinguals have their fingers in two pies—or is it their toes? An electrophysiological investigation of semantic access in bilinguals. *Frontiers in Psychology, 3,* 52–57.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*, 281-300.

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22*(5), 354–367.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163–182.

LaCelle-Peterson, M., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55–76.

Lee, J. F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition, 8,* 201–12.

Lee, K. Q. (2015a). GEPT KIDS Model Test. Taipei, Taiwan: Learning Publishing.

Lee, K. Q. (2015b). GEPT KIDS Reading Comprehension Test. Taipei, Taiwan: Learning Publishing.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–386.

MOE (Taiwan Ministry of Education). (2001). Elementary school 300 fundamental vocabularies. Retrieved from http://www.ntcu.edu.tw/smya/970526.pdf

MOE (Taiwan Ministry of Education). (2008). Junior high school and elementary school 1200 most fundamental vocabularies. Retrieved from https://goo.gl/WqjelB

Otheguy, R., García, O., & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review, 6*(3), 281–307.

Pearson, P. D., & Johnson, D. D. (1978). *Teaching reading comprehension.* New York, NY: Holt, Rinehart, & Winston.

Rahimi, M. (2007). L2 reading comprehension test in the Persian context: Language of presentation as a test method facet. *Reading Matrix: An International Online Journal, 7*(1).

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors.

*Psychological Methods, 13*(1), 19–30.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*(2), 147–170.

Short, D. (1993). Assessing integrated language and content instruction. *TESOL Quarterly, 27*(4), 627–656.

Snow, M. A. (1998). Trends and issues in content-based instruction. *Annual Review of Applied Linguistics, 18*, 243–267.

The National Center for Fair and Open Testing. (2005). Assessment problems of ELL students under NCLB: Problems and solutions. Retrieved from http://www.fairtest.org/sites/default/files/NCLB_and_assessing_bilingual_students.pdf

Vacca, R. T., & Vacca, J. L. (1999). *Content area reading: literacy and learning across the curriculum* (6th ed.). New York, NY: Longman.

Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing, 25*(4), 521-551.

Yun, J. H., Lee, M., & Park, Y. H. (2012). The influence of the language of directions, questions, and choices in practice CSAT on learners' test results. *Korean Journal of Applied Linguistics, 28*(1), 59–85.

## Acknowledgements

I am pleased to thank Sandra, Hardy, and Dr. Tzeng for their kind assistance and guidance in this project. In addition, I am deeply grateful to the reviewers for their comments and encouragement. It was truly a pleasure working with and learning from all of you.

*Corresponding author: curtis2020@gmail.com