Mizumoto, A., Ikeda, M., & Takeuchi, O. (2016). A comparison of cognitive processing during cloze and multiple-choice reading tests using brain activation. *ARELE (Annual Review of English Language Education in Japan), 27*, 65–80.

A comparison of cognitive processing during cloze and multiple-choice reading tests using brain activation

Atsushi Mizumoto
Kansai University
Maiko Ikeda
Kansai University
Osamu Takeuchi
Kansai University

Abstract

This study compares cognitive processing and cerebral activation during a cloze test to a multiple-choice reading test. Data were obtained through an innovative neuroimaging technique (near-infrared spectroscopy, NIRS) and stimulated recall interviews. Fifteen Japanese EFL (English as a foreign language) learners participated. Greater brain activation was observed in the cloze condition than in the multiple-choice condition. Individual variation in degree of cerebral activation was also found and further examined by referencing the stimulated recall interviews. Pedagogical and research implications are provided, especially emphasizing that practitioners and researchers should exercise caution and informed judgment when they use cloze tests.

1. Introduction

According to Tremblay (2011), about 30% of the 53 research articles published between 2000 to 2008 in *Second Language Research* and *Studies in Second Language Acquisition* that reported using any type of independent tests utilized a cloze test (or a C-test) for measuring an individual's second-language (L2) proficiency. The use of cloze tests as a measure of general language proficiency is so common that most introductory books on second language acquisition (SLA), teaching English as a second or foreign language (TESOL/TEFL), and language testing include descriptions and suggestions of cloze tests (e.g., Fulcher & Davidson, 2007; Loewen & Reinders, 2011).

The cloze procedure, originally created as a readability measure of texts (Taylor, 1953), removes every n-th word (e.g., every seventh word) from the original text. When such processed text with the cloze procedure is used as a "cloze test," the test takers are required to fill in the missing words with appropriate words. The two variants of the cloze test, the C-test (Klein-Braley

& Raatz, 1984) and the rational cloze test (Bachman, 1985), are also widely accepted measurement tools that have made revisions on the weaknesses of the original cloze procedure.

In the field of language testing, the research on cloze tests has a long and winding history. Cloze research goes back almost half a century, reaching a peak in the 1970s and 1980s. It would not go too far to suggest that the early work in language testing as a field developed on the foundation of cloze research (Lazaraton, 2010). It is therefore no wonder that the cloze test is one of the all-time most researched topics in language testing.

The cloze test has also been used as a placement test (O'Toole & King, 2011) and a yardstick used to measure gains in language learning and teaching (Ross, 1998). Although similar to a cloze test, the gap-filling test, in which a test writer decides to delete certain words on the basis of a testing objective, often with multiple-choice options, is frequently used as well (see Alderson, 2000 for reasons to distinguish "cloze" and "gap-filling" tests). Both cloze and gap-filling formats are also used as effective vocabulary-learning exercises.

As described above, the cloze is a firmly established testing and learning tool in the field of SLA and foreign or second language teaching and learning. In the following session, we will briefly review the past research findings on the cloze tests by focusing on the specific aspect of "what the cloze test measures."

1.1 Debates over What the Cloze Test Measures

When discussing the development of cloze-test research, it is necessary to understand the context of the evolution of language testing as a research field (for an excellent historical review of language testing by McNamara, 2013). The early period of modern language testing placed great emphasis on objectively testing discrete components of language (i.e., vocabulary, grammar, and phonology) with multiple-choice tests (Lado, 1961). Under the psychometric traditions of the time, such an approach to language testing was considered scientific and appropriate. A breakthrough emerged, however, when Carroll (1972, originally published in 1961) proclaimed the need to test integrated skills in language performance, rather than focusing only on individual components. Further, Oller (1979) suggested that cloze tests would be a good candidate for such "integrative tests," as opposed to discrete-point tests, because the test format requires test takers to integrate various areas of language knowledge. "Integrative" tests should not be confused with "integrated" tests in which, for example, test takers write based on reading some text, such as items in the TOEFL iBT (Alderson, 2009). Oller further claimed that performance on cloze tests "would be predictive of performance on others, because what was being tapped was an integrative performance capacity, something not captured in discrete-point tests" (McNamara, 2013, p. 342), and therefore language proficiency is a unitary concept, not a divisible one. This was called the "unitary competence hypothesis." Although Oller (1983) later abandoned the strongest form of the unitary competence hypothesis (see also Purpula, 2010 for details), the cloze test nonetheless

retained its status as a language-proficiency measurement because it still correlates well with a variety of linguistic skills (e.g., Kobayashi, 2002b; Weir, 1990).

During the 1970s and 1980s, a great deal of research was devoted to identifying exactly what it is that cloze tests measure. Some researchers argued that cloze tests measure only a sentence-level understanding of the text and thus only assesses lower-order skills (e.g., Alderson, 1979). Others have maintained that cloze tests can measure higher-order skills (e.g. Sasaki, 2000). After decades of research, it is now generally accepted that although the types of text, passage difficulty, number of deletions, scoring methods, and proficiency levels of test-takers can have considerable effects on the results, cloze tests measure "abilities including syntactic or grammatical knowledge and both lower-level (e.g., clausal and sentential) and higher-level (e.g., intersentential and textual) reading comprehension abilities" (Yamashita, 2003, p. 268). Brown (2013, p. 9) further suggested that, depending on one's proficiency, cloze tests may tap different abilities. For low-level learners who can only handle sentence-level grammar or vocabulary, test items discriminate learners at that level. For advanced-level learners, cloze tests measure higher-level reading-comprehension abilities because they can deal with both sentence-level grammar and across-sentence cohesion, coherence, and pragmatics. Brown (2013) states, "The potential for both sentential and intersentential items must exist in most passages because that potential exists in the written language" (p. 22).

What cloze tests measure in general is known as "pragmatic expectancy grammar," which predicts the kind of information that will come next given a particular context (Oller, 1979). Cloze tests economically measure a test taker's overall language ability based on the premise that, according to Hughes (2003), "performance in one skill is usually a reasonable predictor of performance in another . . . On some occasions, of course, we may be wrong in our prediction, but usually we will be right." Hughes also notes that "despite their differences, speaking and writing share a great many features, most obviously elements of grammar and vocabulary. It is this sharing of features that allows us to measure overall ability economically" (pp. 186–187). Using cloze tests to measure other skills is so prevalent in the literature that some researchers claim that cloze tests can be a valid measure of reading comprehension (Gellert & Elbro, 2012). Others are more cautious regarding their use as a reading measure. Grabe (2009) warns that cloze tests "are not automatically valid assessments of reading abilities, particularly when students are expected to write in the missing words. Such tests become production measures and are not appropriate for L2 reading assessment" (p. 359). However, considering the overarching latent traits measured by cloze tests (e.g., grammar and vocabulary), it makes intuitive sense that cloze tests, when prepared carefully, have relatively high correlations with other skills, especially reading, and overall proficiency.

1.2 Case *Clozed*? The Possibility of Physiological Data

A review of cloze literature obviously shows that the literature is so vast and comprehensive to date that there no longer remains an uncharted territory in the cloze research endeavor. However, there is one aspect of data obtained from the cloze test takers that have not yet been considered—a physiological data.

Recent years have seen increased interest in the cognitive processing involved in language testing (Bax, 2013; Rupp, Feme, & Choi, 2006). This has been largely motivated by the notion of "cognitive validity," which is the extent to which test tasks and responses properly and correctly reflect underlying cognitive processing (Alderson, 2000; Field, 2011). This growing interest in the cognitive aspects of test tasks can also be found in cognitive diagnostic assessment (CDA), which has gained recognition for providing more detailed diagnostic information regarding test takers' performance (e.g., Lee & Sawaki, 2009).

In order to take a closer look at the cognitive processes underlying test tasks, some researchers have turned to physiological data, often coupled with innovative measurement technology such as eye-tracking (Bax, 2013) and functional magnetic resonance imaging (fMRI) (Jeong et al., 2011). Physiological data has yet to be considered in research on cloze tests. In previous studies on cloze tests, elicitation methods such as think-aloud protocols have been utilized to investigate cognitive processes (e.g., Storey, 1997). An event-related brain potential (ERP) study by DeLong, Urbach, and Kutas (2005), although not particularly focused on language testing, reported that readers used the words in a sentence to estimate the relative likelihoods of upcoming words (i.e., an anticipatory cognitive process). However, no study to date has examined brain activation during cognitive processing in a cloze test. In this study, we used brain-imaging methods similar to our previous studies of cerebral activation during reading-aloud activities (Takeuchi, Ikeda, & Mizumoto, 2012a) and the use of reading strategies (Takeuchi, Ikeda, & Mizumoto, 2012b). These methods provide a valid and reliable physiological measure of underlying brain activity.

Cloze tests require readers "to construct meaning through greater levels of language awareness than in normal reading" (Raymond, 1988). Considering this and the productive nature of cloze tests (Grabe, 2009; Kobayashi, 2002a), we hypothesized that cloze tests may require more cognitive resources than standard reading-comprehension tests with multiple-choice questions (MCQ), and predicted that answering cloze-test questions would elicit greater cerebral activation than answering MCQ.

2. Method

2.1 Participants

Fifteen healthy right-handed volunteers (7 females and 8 males) participated in the study. All the participants were right-handed individuals because handedness has an effect on the

functioning of the brain. They were intermediate-to-advanced Japanese EFL learners with mean paper-based TOEFL scores of 553.27 (*SD* = 54.32). Ages ranged from 18 to 47 years (*M* = 25.00, *SD* = 8.16). We recruited proficient EFL learners in the current study so that they could complete the cloze and reading tasks without much difficulty. Eight participants were undergraduate students (English majors), five were graduate students (in a Teaching English to Speakers of Other Languages [TESOL] Master of Arts [MA] program), and two were university lecturers in English. We obtained written informed consent and personal information from participants after providing a complete description of the experimental procedures and the purpose of the study as well as the ethical responsibility of the researchers. Each participant received a bookstore gift certificate for 1,000 Japanese Yen.

2.2 Instruments

We used a brain imaging technique called Near-Infrared Spectroscopy (NIRS), also known as "optical topography," which is a real-time, non-invasive brain-imaging technique requiring less participant restraint than fMRI. NIRS uses near-infrared light to estimate changes in cerebral blood volume and oxygen saturation, which are indicators of brain activity. The NIRS is used for the research of recognition, language processing, and thinking processes due to its non-invasive and real-time natures. A number of studies have utilized NIRS as a satisfactory method to measure brain activity (e.g., Ehlis, Herrmann, Wagener, & Fallgatter, 2005; Tsujimoto, Yamamoto, Kawaguchi, Koizumi, & Sawaguchi, 2004). We used the ETG-4000 Optical Topography System (Hitachi Medical Co., Japan) with a 52-channel array of optodes, which measured activation in most areas of the prefrontal cortex, corresponding to the functioning of working memory (e.g., Curtis & D'Esposito, 2003).

Two reading passages were prepared: one using multiple-choice questions and one using a cloze format. Both tests were based on the 2nd Grade of the EIKEN Test in Practical English Proficiency, produced by the Eiken Foundation of Japan. We chose passages from the same level (i.e., 2nd grade) so that they would be of a similar difficulty. For MCQ, we used a passage and questions taken directly from the EIKEN Test. For the cloze format, we employed the most standard procedure, deleting every seventh word of a passage (Schmitt, 2000, p. 152) but leaving the first sentence intact (http://mizumot.com/files/ARELE2016Appendix.pdf).

The difficulty of the passages (Table 1) was checked with (a) a readability index (Flesch-Kincaid Grade Levels), (b) text easability scores obtained from Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) and (c) word levels obtained using VocabProfile (Cobb, n.d.). Although the passage for multiple-choice questions (i.e., "Dogs That Count") was easier in terms of readability and text easability, we asked participants during the interview session following the experiment and confirmed that passages for cloze and multiple-choice question tasks were nearly the same in perceived difficulty and that neither was too demanding. In addition, topic familiarity

was taken into consideration when selecting text passages, and in the interview session, we confirmed that topic familiarity did not pose problems for text comprehension.

We did not include a "normal" reading task, in which participants read the text without any particular goal or conscious application of reading strategies, because (a) it was not directly related to the hypothesis being tested and (b) a previous study (Takeuchi et al., 2012b), has already shown baseline activation due to normal reading.

Table 1
Summary of Task Reading Passages

Task		Multiple-choice (MCQ)	Cloze
Title of the Passage		*Dogs That Count*	*False Memories*
Passage Length (Words)		367	362
Readability (Flesch-Kincaid Grade Level)		8.9	10.4
Text Easability (Percent)	Narrativity	51.20	53.19
	Syntactic Simplicity	58.32	33.36
	Word Concreteness	77.94	64.43
	Referential Cohesion	60.64	46.81
	Deep Cohesion	93.32	64.80
Word Levels (Percent)	K1 Words (1–1000)	87.03	87.81
	K2 Words (1001–2000)	5.95	4.99
	AWL Words (Academic)	2.16	3.88
	Off-List Words	4.86	3.32

Note. See Appendix for the actual passages. Coh-Metrix (http://cohmetrix.com/) was used to compute text *easability*. Greater easability scores indicate easier text. VocabProfile (http://www.lextutor.ca/vp/eng/) was used to calculate Word Levels.

2.3 Procedures

All procedures followed the principles of the Declaration of Helsinki (World Medical Association, 2008). The experiment was conducted in a quiet room. Each participant sat in a chair with a task sheet attached to an adjustable plastic holder placed on the desk in front of him or her. During tasks, each participant was asked to read the passages and answer the questions with his or her pen, pointing at the place where his or her eyes were fixated. This enabled us to ascertain the approximate places where each participant was reading via videotaping with two cameras. This location information, although rough, was used to relate changes in NIRS measurements to participants' reading behavior, and it was used in the subsequent interview sessions.

During the experiment, three tasks were presented to each participant for 120 s each: (1) multiple-choice questions, (2) a cloze-format passage, and (3) a writing-down task in which

participants were asked to write random English words. The writing task was included to check whether the effects of writing would alter the degree of cerebral activation. It was used as a baseline (i.e., control) in this experiment. The order of MCQ and cloze tasks was counterbalanced across participants. A 60-second rest period was included after each task. During these breaks, participants were instructed to relax and silently read a piece of paper with the letters of the Latin alphabet (A to Z) for the purpose of canceling out the effects of the preceding task. This is a standard procedure in brain-imaging studies using NIRS. Prior to the experiment, participants were provided with both a detailed explanation of each task and an opportunity to practice the tasks with sample passages.

After each participant completed all tasks, a stimulated recall interview was conducted to complement the NIRS data. Stimulated recall is a method to collect learners' insights by providing them with a stimulus, such as an audio or video recording, and asking them to recall thoughts they had while completing a specific task (Gass & Mackey, 2000). In our stimulated recall interview, we showed each participant the tasks again along with the video clip of him or herself working on each task. Graphical representations of changes in their blood hemoglobin concentrations were also shown, synchronized with one of the video cameras. This enabled us to pinpoint the place during the task where a rise or decrease of cerebral activation occurred with participant-reported indications of what he or she was actually thinking at that time. The interview session was recorded with an IC recorder. The entire experiment took approximately 60 min for each participant, including instructions and interviews.

2.4 Data Analyses

We analyzed the relative changes in oxy-Hb (the unit of measurement is millimolar × millimeter, mM-mm) during the tasks. We chose oxy-Hb over deoxy-Hb, because prior NIRS studies have indicated that concentration of oxy-Hb is a clearer and more reliable indicator of brain activity than deoxy-Hb (e.g., Tsujii, Yamamoto, Ohira, Saito, & Watanabe, 2007). To obtain the relative changes in hemoglobin concentration precisely, we used a method called "integral analysis," which applies a linear-fit function for a baseline correction and uses rest periods for pre-task and post-task baselines. We calculated the average concentration of oxy-Hb for each participant during each task.

We employed a repeated-measures design with type of task (control, multiple-choice, or cloze) as an independent factor and concentration of oxy-Hb as a dependent variable. To test whether cloze-tests elicited a greater degree of cerebral activation than MCQ, we applied a multilevel model (Hox, 2002) with restricted maximum likelihood estimation. Models of this sort are an extension of regression models that can handle non-independent, repeated-measures data more properly (Field, Miles, & Field, 2012). In recent years, many researchers have been advocating the use of such models over conventional ANOVAs in the field of applied linguistics and language testing (e.g., Barkaoui, 2013; Kozaki & Ross, 2011). In our multilevel model, we set

orthogonal contrasts to test our predicted comparisons between (a) baseline and both multiple-choice and cloze tasks, and (b) the multiple-choice task and the cloze task. The threshold for statistical significance was set to .05 for all analyses. R version 3.1.2 (R Core Team, 2014) was used for all quantitative analyses. For transparency, data and R codes used in this study are available online (http://mizumot.com/files/ARELE2016.html); therefore, readers can check the raw data, except the personal information, and replicate the analytical procedures.

The qualitative data obtained from stimulated recall interviews were transcribed, coded, and used to corroborate the findings of the quantitative analyses.

3. Results

The descriptive statistics of changes in oxy-Hb concentrations for each task are presented in Table 2. The mean oxy-Hb concentration scores of the target tasks (both multiple-choice questions and cloze format) were higher than for the baseline task (writing down random English words); however, the variance was large. A possible reason for this may be individual differences in physiological data (Morishita & Osaka, 2008). The multilevel analysis indicated that type of task had a statistically significant effect on blood hemoglobin concentrations, $t(29) = 2.917$, $p = .007$, effect size r [95% CI] = .476 [.140, .714]. The orthogonal contrasts revealed that the average concentration of oxy-Hb for the two target tasks (multiple-choice and cloze combined) was statistically greater than the baseline task, $b = 0.036$, $t(28) = 2.712$, $p = .011$, effect size r [95% CI] = .456 [.115, .701]. Concentration of oxy-Hb was also significantly greater for close tasks than for MCQ, $b = 0.050$, $t(28) = 2.139$, $p = .041$, effect size r [95% CI] = .375 [.012, .648]. These results support the research hypothesis of the current study that answering a cloze format elicits higher degrees of cerebral activation than answering MCQ (multiple-choice questions) does.

Table 2
Descriptive Statistics of the Oxy-Hb Concentration during Tasks

Tasks	Mean	*SD*	min	max	*SE*	95% CI	
						Lower	Upper
Baseline	0.012	0.114	-0.208	0.223	0.029	-0.051	0.075
MCQ	0.071	0.133	-0.181	0.285	0.034	-0.003	0.145
Cloze	0.170	0.168	-0.006	0.521	0.043	0.077	0.263

Note. $N = 15$; MCQ: multiple-choice questions. The unit of measurement is millimolar × millimeter (mM-mm).

Figure 1 presents the mean changes in concentrations of oxy-Hb with 95% confidence intervals (CIs) for each task for all 15 participants. Cognitive processing during the tasks varied greatly between individuals, reflecting the fact that individual differences played a role in the processing of each task. Although the overall means of the cloze task were greater than the multiple-choice task and the baseline writing-down task, for some participants, the concentration

of oxy-Hb was greater for the multiple-choice than the cloze task. We focused on these irregularities in our qualitative stimulated recall excerpts.

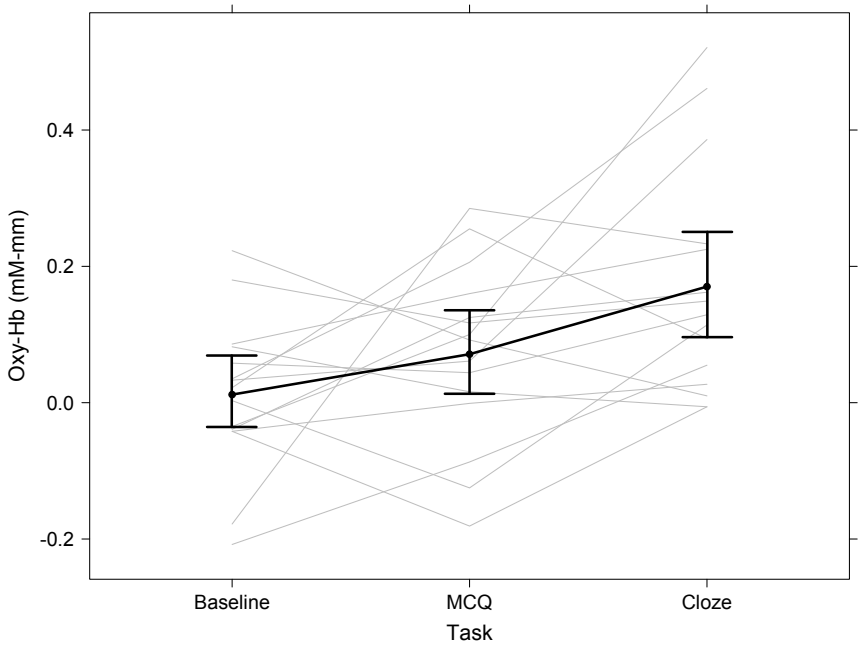

Figure 1. Trajectories of the 15 participants' concentration of oxy-Hb during each task (thin gray lines in the figure). The means for the three tasks are overlaid and represented in points with the thick line. Error bars show 95% confidence intervals (CIs).

The following excerpts (translated to English) were drawn from those participants showing greater cerebral activation during the cloze task compared to the multiple-choice task:

> While I was answering the cloze items, I thought it was really difficult to answer those items. I first thought about appropriate set phrases such as "come up with," and then considered the grammar structure. I constantly paid attention to the meanings so that my answer would make sense. [Cloze > MCQ: Participant 2]

> Depending on the blank in the cloze task, I used knowledge of phrases or grammar. Answering cloze questions was more difficult for me because MCQ required me to only check the content of the passage. On the other hand, cloze questions involved so much more thinking especially when there were a few possible answers for the blank. [Cloze > MCQ: Participant 6]

> When answering the cloze items, I first focused on the meanings when filling in the blanks. If necessary, I relied on grammar knowledge. For the MCQ task, I read sentence by sentence without paying much attention to what I was doing while reading. [Cloze > MCQ: Participant 12]

These excerpts highlight the fact that the participants conducted both analytical and meaning-focused cognitive processing when answering the cloze items, which presumably resulted in higher blood hemoglobin concentrations. It should be noted that, as Participants 6 and 12 remarked, these participants may not have employed specific reading strategies during the MCQ task.

In contrast, the following excerpts are from those who exhibited greater cerebral activation during the MCQ task:

> [On the MCQ task], while I was reading the first paragraph, I also tried to summarize it. I did so because normally with these type of MCQs, I first check the [four] options. Because of that constraint, I thought it would be better to summarize the passage before checking the options. [Cloze < MCQ: Participant 15]

> For the cloze task, I was analyzing the grammar structure first and then paid attention to the contextual meanings. If using grammar knowledge didn't work, for example, for filling in some content words, I thought about the context. For the MCQ task, after reading the first paragraph, I checked the stem [not the options] of the first question and went back to the first paragraph to check the content. Then I chose the most appropriate option by summarizing the main idea. [Cloze < MCQ: Participant 9]

From these comments, it seems that participant exhibiting greater activation during the multiple-choice task employed particular types of reading strategies. It has been proposed (Macaro, 2006) and validated (Takeuchi et al., 2012b) that one's reading strategy can elicit the perceiving, holding, processing, and encoding functions of working memory, reading strategy use could account for greater cognitive processing while answering multiple-choice questions than cloze items.

4. Discussion

We tested the hypothesis that cloze tasks require greater cognitive processing than multiple-choice tasks in tests of language learning using NIRS brain imaging. Overall, brain-imaging results supported this hypothesis, with greater mean cerebral activation for cloze tasks than for multiple-choice and control tasks. However, individual variation in the degree of cerebral activation was high, and some participants were inconsistent with the hypothesis.

Stimulated recall interviews revealed that those with a greater concentration of oxy-Hb for the multiple-choice task than the cloze task may have been using specific reading strategies.

Comparing these findings with the literature, our results highlight one clear aspect of cloze tests: They are just one testing technique, like MCQ (Alderson, 1979), and using cloze tasks does not guarantee cognitively demanding processing. Summarizing his 25 years of cloze-testing research, J. D. Brown (2013) added:

> yes, cloze tests are just another technique for creating contextualized test items, but it is a technique that is not very efficient in terms of developing items at the appropriate level of difficulty that discriminate well in second language populations . . . In other words, a cloze test that is not tailored is just an inefficient collection of unpiloted items. Do you really want to administer such a raw test to your students when you are making the important sorts of high-stakes proficiency and placement decisions you make with norm-referenced tests? (pp. 26–27)

Thus, any cloze test can function poorly and the results are unpredictable unless the test has been validated (Alderson in his interview with Brunfaut, 2014). One cannot simply administer a cloze test, naively believing that this format will always measure integrative language ability. Task items need to be pilot tested, analyzed, and revised to confirm that a particular cloze test is valid for the target test takers. These requirements will be true of any language test, yet are too often neglected when it comes to cloze tests (Brown, 2013). In this sense, using modified cloze variants such as C-tests (Klein-Braley & Raatz, 1984) and rational cloze tests (Bachman, 1985), which have been invented to overcome some shortcomings of the cloze test, may be a better option when, due to some constraints, pilot testing, item analyses, and item revisions are not feasible.

Using cloze tests that have not been validated also introduces problems with research replication. Although cloze tests are often used for proficiency-assessment methods in the field of applied linguistics (Tremblay, 2011), it is often the case that in the original study a detailed description of the cloze test is not provided, with only the statement "proficiency was measured with a cloze test." We agree with Fitzpatrick (2012), who stated, "In order to conduct any kind of replication the researcher must design what he or she hopes will be an equivalent test, and make decisions about text selection, length of text, number and frequency of gaps, and so on" (p. 159). As can be expected, randomly created, idiosyncratic "cloze tests" do not yield standardized, comparable, or reliable results across studies because they may measure different aspects of L2 ability, depending on the characteristics of the test takers (Brown, 2013). The results of the current study support such complicating individual variation in test-taking strategies. Thus, for replication to be made possible, cloze test should be made accessible to anyone intending to measure the proficiency of L2 learners and compare results across studies (e.g., see exemplary work by Tremblay, 2011). It should also be noted that although meta-analyses have been on the rise in

recent years in applied linguistics (Oswald & Plonsky, 2010) and in language testing (e.g., In'nami & Koizumi, 2009), these comparisons should be interpreted with extreme caution when they consider incomparable cloze results in the primary literature (Watanabe & Koyama, 2008). For instance, an "intermediate-level proficiency" measured with an unstandardized cloze test in one study may not indicate the same level in another study with a different cloze instrument.

Although the present study sheds new light on the cognitive processing of cloze tests with an innovative brain-imaging technique, near-infrared spectroscopy, it does have two important limitations. The first concerns the possible effect of differences between passages in the cloze and multiple-choice tasks. Although text selection was based on text difficulty and topic familiarity, it was impossible to have two identical passages. Small differences between the passages may have affected the results of this study. The second limitation is that, following the convention of studies using near-infrared spectroscopy, we calculated the average concentration of oxy-Hb during each task. However, as repeatedly pointed out in the literature (Bachman, 1985), each cloze item requires different types of contextual clues to complete (i.e., within clause, across clause, within sentence, across sentence, within text, and extra-textual). Each cloze item, therefore, does not carry the same amount or kind of information (Todd & Gu, 2007, p. 17). Furthermore, it is clear in the current study that some participants that showed a greater degree of cerebral activation during the multiple-choice task used some elaborate cognitive strategies. It is thus conceivable that more detailed pictures of cognitive processing during test-taking would be available upon analyzing each test item in terms of anticipated cognitive processing (e.g., Bax, 2013; Rupp, Feme, & Choi, 2006). A brain-imaging study including item-by-item analysis should be implemented in future studies. Even with these limitations, the current study provides physiological evidence that cloze testing elicits a high degree of cerebral activation. At the same time, it demonstrated that cognitive processing during the answering of test items can be investigated with brain-imaging techniques. We believe this approach will lead to a better and deeper understanding of the cognitive processing of test takers and language learners.

With this research, it was not our intention to rekindle old debates over cloze tests, which "were closed decades ago" (Spolsky, 2010, p. 451). Rather, in hindsight, we realized that decades-old findings of the cloze literature could be still applicable even when cloze tests were investigated with cutting-edge technology (i.e., NIRS). Although we now live in the 21st-century world of communicative language teaching and testing, with much emphasis on "integrated performance on whole tasks" (McNamara, 2013), cloze tests continue to be used for research and pedagogical purposes. Cloze research findings are therefore relevant and important even today (Brunfaut, 2014), and cloze tests must still be used and interpreted properly.

5. Conclusion

The findings of the present study suggest that on average cloze-format language tests elicit greater cognitive processing than multiple-choice tests, as measured by NIRS blood hemoglobin concentration. Individual variation in the degree of cerebral activation was also found and further examined by referring to the stimulated recall interviews. As a result, we attribute this individual variation the fact that different test takers may utilize different cognitive strategies, and that (as largely agreed in the literature) employing a cloze procedure does not always automatically guarantee deeper cognitive processing. One important implication of the current research is that practitioners and researchers should exercise extreme caution when using cloze tests that have not been rigorously validated.

Acknowledgments

This study was financially supported by JSPS KAKENHI (Grant Numbers 20520540 and 80454768).

References

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, *13*, 219–227. doi:10.2307/3586211

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

Alderson, J. C. (2009). Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing*, *26*, 621–631. doi:10.1177/0265532209346371

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*, 535–556. doi:10.2307/3586277

Barkaoui, K. (2013). Using multilevel modeling in language assessment research: A conceptual introduction. *Language Assessment Quarterly*, *10*, 241–273. doi:10.1080/15434303.2013.769546

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*, 441–465. doi:10.1177/0265532212473244

Brown, J. D. (2013). My twenty-five years of cloze testing research: So what? *International Journal of Language Studies*, *7*, 1–32. Retrieved from https://www.academia.edu/8130941/My_twenty-five_years_of_cloze_testing_research_So_what

Brunfaut, T. (2014). A Lifetime of language testing: An interview with J. Charles Alderson. *Language Assessment Quarterly*, *11*, 103–119. doi:10.1080/15434303.2013.869818

Carroll, J. B. (1972). Fundamental considerations in testing for English language proficiency in foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings* (2nd ed., pp. 313–321). New York, NY: McGraw-Hill.

Cobb, T. (n.d.). VocabProfile. Retrieved from http://www.lextutor.ca/vp/

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*, 415–423. doi:10.1016/S1364-6613(03)00197-9

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117–1121. doi:10.1038/nn1504

Ehlis, A. C., Herrmann, M. J., Wagener, A., & Fallgatter, A. J. (2005). Multi-channel near-infrared spectroscopy detects specific inferior-frontal activation during incongruent Stroop trials. *Biological Psychology*, *69*, 315–331. doi:10.1016/j.biopsycho.2004.09.003

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London, UK: Sage.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking, Studies in Language Testing 30* (pp. 65–111). Cambridge University Press and Cambridge ESOL.

Fitzpatrick, T. (2012). Conducting replication studies: Lessons from a graduate program. In G. Porte (Ed.), *Replication research in applied linguistics* (pp. 151–170). Cambridge University Press.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge.

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

Gellert, A. S., & Elbro, C. (2012). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment*, *31*, 16–28. doi:10.1177/0734282912451971

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

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

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

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*, 219–244. doi:10.1177/0265532208101006

Jeong, H., Hashizume, H., Sugiura, M., Sassa, Y., Yokoyama, S., Shiozaki, S., & Kawashima, R. (2011). Testing second language oral proficiency in direct and semidirect settings: A social-cognitive neuroscience perspective. *Language Learning*, *61*, 675–699. doi:10.1111/j.1467-9922.2011.00635.x

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-Test. *Language Testing*, *1*, 134–146. doi:10.1177/026553228400100202

Kobayashi, M. (2002a). Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *Modern Language Journal*, *86*, 571–586. doi:10.1111/1540-4781.00162

Kobayashi, M. (2002b). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, *19*, 193–220. doi:10.1191/0265532202lt227oa

Kozaki, Y., & Ross, S. J. (2011). Contextual dynamics in foreign language learning motivation. *Language Learning*, *61*, 1328–1354. doi:10.1111/j.1467-9922.2011.00638.x

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London, UK: Longmans, Green.

Lazaraton, A. (2010). From cloze to consequences and beyond: An interview with Elana Shohamy. *Language Assessment Quarterly*, *7*, 255–279. doi:10.1080/15434301003792815

Lee, Y.-W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, *6*, 172–189. doi:10.1080/15434300902985108

Loewen, S., & Reinders, H. (2011). *Key concepts in second language acquisition*. Basingstoke, UK: Palgrave Macmillan.

Macaro, E. (2006). Strategies for language learning and for language use: Revising the theoretical framework. *Modern Language Journal*, *90*, 320–337.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

McNamara, T. (2013). Language testing: History, validity, policy. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology. APA handbooks in psychology.* (pp. 341–352). Washington, DC: American Psychological Association. doi:10.1037/14047-021

Morishita, M., & Osaka, N. (2008). Information processing and working memory capacity in linguistic working memory tasks. In N. Osaka (Ed.), *Neural Correlates of Working Memory* (pp. 123–158). Kyoto, Japan: Kyoto University Press.

O'Toole, J., & King, R. (2011). The deceptive mean: Conceptual scoring of cloze entries differentially advantages more able readers. *Language Testing*, *28*, 127–144. doi:10.1177/0265532210375687

Oller, J. W. (1979). *Language tests at school*. London, UK: Longman.

Oller, J. W. (1983). Response to Vollmer: "g", what is it? In A. Hughes & D. Porter (Eds.), *Current developments in language testing* (pp. 35–37). London, UK: Academic Press.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. doi:10.1017/S0267190510000115

Purpula, J. E. (2010). Assessing communicative language ability: Models and their components. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., pp. 53–68). New York, NY: Springer.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. Retrieved from http://www.r-project.org/

Raymond, P. (1988). Cloze procedure in the teaching of reading. *TESL Canada Journal*, *6*, 91–97. Retrieved from http://www.teslcanadajournal.ca/index.php/tesl/article/view/544/375

Ross, S. (1998). *Measuring gain in language programs: Theory and research*. Sydney, Australia: National Centre for English Language Teaching and Research.

Rupp, A. A., Feme, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*, 441–474. doi:10.1191/0265532206lt337oa

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing*, *17*, 85–114. doi:10.1191/026553200671343210

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.

Spolsky, B. (2010). Language assessment in historical and future perspective. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., pp. 445–454). New York, NY: Springer.

Storey, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse cloze test. *Language Testing*, *14*, 214–231. doi:10.1177/026553229701400205

Takeuchi, O., Ikeda, M., & Mizumoto, A. (2012a). Reading aloud activity in L2 and cerebral activation. *RELC Journal, 43,* 151–167. doi:10.1177/0033688212450496

Takeuchi, O., Ikeda, M., & Mizumoto, A. (2012b). The cerebral basis for language learner strategies: A near-infrared spectroscopy study. *Reading in a Foreign Language, 24,* 136–157. Retrieved from http://nflrc.hawaii.edu/rfl/October2012/articles/takeuchi.pdf

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*, 415–433.

Todd, L., & Gu, P. (2007). Rational cloze tests as a placement measure for EAL learners. *The TESOLANZ Journal*, *15*, 16–29. Retrieved from http://www.tesolanz.org.nz/includes/download.aspx?ID=101642

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research. *Studies in Second Language Acquisition*, *33*, 339–372. doi:10.1017/S0272263111000015

Tsujii, T., Yamamoto, E., Ohira, T., Saito, N., & Watanabe, S. (2007). Effects of sedative and non-sedative H1 antagonists on cognitive tasks: Behavioral and near-infrared spectroscopy (NIRS) examinations. *Psychopharmacology*, *194*, 83–91. doi:10.1007/s00213-007-0814-z

Tsujimoto, S., Yamamoto, T., Kawaguchi, H., Koizumi, H., & Sawaguchi, T. (2004). Prefrontal cortical activation associated with working memory in adults and preschool children: An event-related optical topography study. *Cerebral Cortex*, *14*, 703–712. doi:10.1093/cercor/bhh030

Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies*, *26*, 103–133. Retrieved from http://www.hawaii.edu/sls/wp-content/uploads/2014/09/Watanabe_Koyama.pdf

Weir, C. J. (1990). *Communicative language testing*. Hertfordshire, UK: Prentice Hall International.

World Medical Association. (2008). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. Retrieved from http://www.wma.net/e/policy/b3.htm

Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, *20*, 267–293. doi:10.1191/0265532203lt257oa